LEARNING AN INTEGRATED HYBRID IMAGE RETRIEVAL SYSTEM

A Thesis
Presented to
The Academic Faculty

by

Yushi Jing

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Computing

Georgia Institute of Technology
May 2012

LEARNING AN INTEGRATED HYBRID IMAGE RETRIEVAL SYSTEM

Approved by:

James M. Rehg, Committee Chair
College of Computing
Georgia Institute of Technology

Irfan Essa
College of Computing
Georgia Institute of Technology

Michele Covell
Google Research
Google Inc.

Kiyoharu Aizawa
Department of Electrical Engineering
Tokyo University

Aaron Bobick
College of Computing
Georgia Institute of Technology

Date Approved: Dec 2011

for dad and gugu

ACKNOWLEDGEMENTS

There are many people who have helped and supported me find my way along a very long and winding path towards the completion of my doctoral thesis.

First, I am fortunate to have Dr. James Rehg at Georgia Tech as my advisor, and Dr. Michele Covell at Google Research who played the role of a co-advisor. Throughout my graduate study, Jim's comprehensive knowledge in the field enabled me to identify the broader picture emerging from the research details. Michele has been a dedicated mentor and supporter who encouraged me to combine visual analysis with large-scale Web retrieval. It is her encouragement and support that kept my spirit high and hope alive.

I like to thank my thesis committee members: Dr. Irfan Essa, Dr. Aaron Bobick and Dr. Kiyoharu Aizawa, for their helpful comments and critiques at various stages of this work. I have known Dr. Essa and Dr. Bobick from my undergraduate years, and I have always enjoyed discussing research topics with them. I also appreciate Dr. Aizawa's insights from the thesis proposal that helped me making faster progress.

I would also like to thank my colleagues at Google Research for the wonderful discussions and collaborations. In particular, I thank Dr. Shumeet Baluja and Dr. Henry Rowley for taking me under their wings at Google and introducing me to the world of large-scale data analysis. I thank my managers Dr. Sergey Ioffe and Jay Yagnik for giving me the freedom to explore. I like to thank Dr. Jingbin Wang and Dr. Chuck Rosenberg for helping me to understand large-scale Web image search. I also thank my fellow students, colleagues and friends Dr. Jianxin Wu, Dr. Howard Zhou, Dr. Charlie Brubaker, Dr. Yi Liu, Tim Tang, Rohan Seth, David Tsai for the discussions, collaborations and support.

Finally I thank my dad who completed his Ph.D. study in medicine at the age of 44, and my aunt Gugu who finished her Ph.D. in Chinese literature at the age of 45, for showing me the power of commitment and perseverance. This thesis is dedicated to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Current Web image search engines, such as Google or Bing Images, adopt a hybrid search approach in which a text-based query (e.g. "apple") is used to retrieve a set of relevant images, which are then refined by the user (e.g. by re-ranking the retrieved images based on similarity to a selected example). This approach makes it possible to use both text information (e.g. the initial query) and image features (e.g. as part of the refinement stage) to identify images which are relevant to the user. One limitation of these current systems is that text and image features are treated as independent components and are often used in a decoupled manner. This work proposes to develop an integrated hybrid search method which leverages the synergies between text and image features. Recently, there has been tremendous progress in the computer vision community in learning models of visual concepts from collections of example images. While impressive performance has been achieved on standardized data sets, scaling these methods so that they are capable of working at web scale remains a significant challenge. This work will develop approaches to visual modelling that can be scaled to address the task of retrieving billions of images on the Web.

Specifically, we propose to address two research issues related to integrated text- and image-based retrieval. First, we will explore whether models of visual concepts which are learned from collections of web images can be utilized to improve the image ranking associated with a text-based query. Second, we will investigate the hypothesis that the click-patterns associated with standard web image search engines can be utilized to learn query-specific image similarity measures that support improved query-refinement performance. We will evaluate our research by constructing a prototype integrated hybrid retrieval system based on the data from 300K real-world image queries. We will conduct user-studies to evaluate the effectiveness of our learned similarity measures and quantify the benefit of our method in real world search tasks such as target search.

CHAPTER I

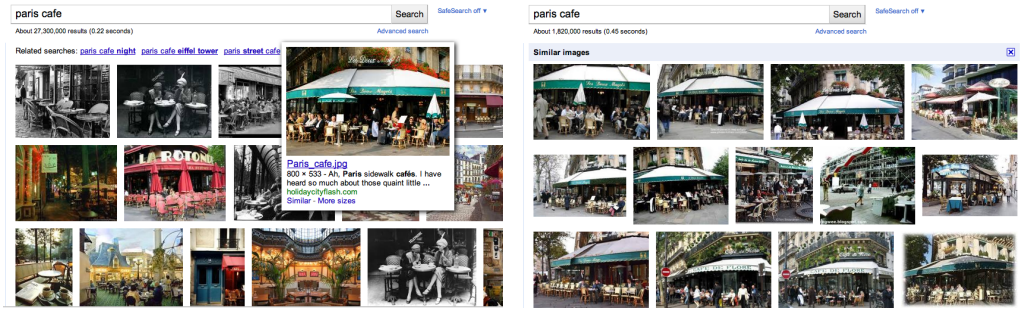
INTRODUCTION

1.1 Hybrid Image Retrieval Systems

Imagine that you just returned from a trip to Paris and want to write about the famous café located near *Saint-Germain-des-Prés*. Although you do not remember its name, you can still remember the distinctive look of its dome-shaped awning as shown in Figure 1. If you kept a photo of this café, then content-based image retrieval systems [28, 96, 23] can be used to find a similar photo on the Web. Otherwise you may use a general text query such as "Paris cafe" to describe the photo and use meta-data based image retrieval systems [12, 11] to retrieve a set of related images. Although images are ranked with respect to relevance scores to the query, in practice it is difficult to know what a user really wants based on a set of keywords, and even more difficult to estimate relevance based on the text meta-data associated with the Web images. As a result, users often need to browse through large set of images before the desired image is found. Our goal is to address this problem by leveraging the availability of large-scale web search data in conjunction with recent methods for learning visual concepts.

Current web image search engines, such as Google or Bing Images, adopt a hybrid search approach in which a text-based query (e.g. "apple") is used to retrieve a set of relevant

Figure 1: A photo of Café les Duex Magots



(a) step 1 (text-based image retrieval) (b) step 2 (content-based image re-ranking)

Figure 2: To find the photo shown in Figure 1, one can use a hybrid Web image retrieval system.

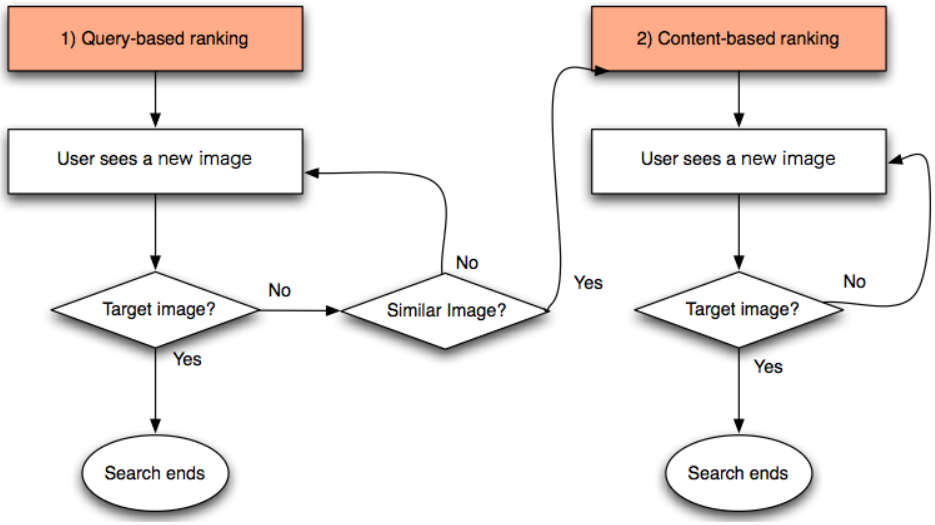


Figure 3: A variation of hybrid image retrieval process

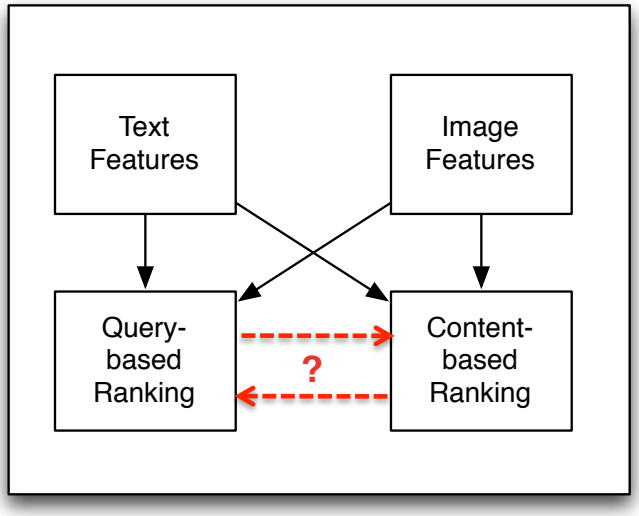


Figure 4: One limitation of these current systems is that text and image features are treated as independent components and are often used in a decoupled manner.

images, which are then refined by the user (e.g. by re-ranking the retrieved images based on similarity to a selected example). An example of such retrieval process is shown in Figure 2. This approach to image retrieval makes it possible to use both text information (e.g. the initial query) and image features (e.g. as part of the refinement stage) to identify images which are relevant to the user. Such *hybrid* image retrieval system [1] is a combination two separate retrieval processes: the first part retrieves images based on matching the text query with the annotations associated with the Web images and the second part computes similarity for images in the search results. The first part is commonly referred as text-based image retrieval and the second part as Query-by-Example (QBE) [105, 75, 49], a variation of content based image retrieval system. The retrieval process of the hybrid system is shown in Figure 3.

One limitation of these current systems is that text and image features are treated as independent components and are often used in a decoupled manner, in the sense that the rank of images produced by one type of retrieval system (e.g. content-based) does not affect the other type (e.g. query-based), as shown in Figure 4. With tremendous engineering and research effort put into improving each component of the Web-scale hybrid image retrieval

[1] It is also referred as composite image retrieval system in [23].

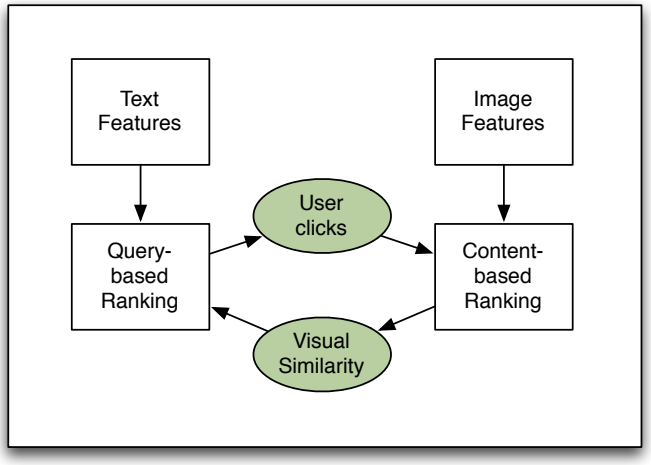


Figure 5: This thesis presents an *integrated* hybrid search method that leverages the synergies between the content- and query-based component of the hybrid image retrieval system.

system [4, 37, 97, 27, 54, 23, 96], it is therefore beneficial to study learning methods that allow improvement in one component of the retrieval system to benefit the other.

This thesis presents an *integrated* hybrid search method that leverages the synergies between the two components, shown in Figure 5. Recently, there has been tremendous progress in the computer vision community in learning models of visual concepts from collections of example images. While impressive performance has been achieved on standardized data sets, scaling these methods so that they are capable of working at web scale remains a significant challenge. This work will develop approaches to visual modelling that can be scaled to address the task of retrieving billions of images on the Web.

An important methodological issue in this research area is how to evaluate and compare image retrieval systems. Although individual components of the system can be evaluated with relevance or precision/recall scores on labelled data sets, a more direct approach is to measure the improvement in user performance (e.g. time-to-completion) on actual image retrieval tasks. In addition to use labelled data-sets to evaluate our proposed approach, we also study image retrieval system on target search [53, 80], a commonly conducted retrieval tasks.

1.2 Thesis Outline

This work studies an integrated hybrid search method that leverages the synergy between text-based and content-based retrieval methods.

Specifically, we study questions that arise when these two components are closely integrated together:

1. *Can we generate improved rankings for text-based image retrieval by measuring the centrality of visual concepts associated with text queries?*

2. *Can we exploit user-click data within a web image search system to automatically learn query-specific image similarity functions?*

3. *Can an integrated hybrid image retrieval system yield quantitative improvements in user performance on retrieval tasks such as target search?*

The answers to these three questions can help us understand whether learning to rank images using image features and users' click patterns (derived from text-based search) can improve users' efficiency in completing retrieval tasks.

This work contains three studies: Chapter 2 studies the hypothesis that for images produced by text-based image retrieval systems, those with higher "centrality" scores (computed from visual features) are perceived to be more relevant to the text query than those with lower scores. We present an efficient approach to compute the Web image centrality scores, and demonstrate that re-ranking image search results with centrality scores can significantly reduce the number of irrelevant search results. Chapter 3 and 4 study the hypothesis that learning query-specific distance functions, with training data derived from the click patterns of text-based search engine users, can improve the estimation of image-to-image similarity.

Chapter 5 studies whether learning to rank using image features and user click patterns, with approaches described in chapter 2 and 3, can indeed improve user performance (e.g. time-to-completion) in completing specific Web image retrieval tasks. The study is conducted by first developing an integrated hybrid image retrieval system that supports the

retrieval of approximately 250 million Web images, and then asking users to perform a set of target-search [96] tasks using such systems.

CHAPTER II

RANKING IMAGES BASED ON VISUAL SIMILARITY

Is there a particular mental image that you would associate with "Mona Lisa?"

We conjecture that if a person is familiar with a concept such as Mona Lisa, he or she is likely to have one or more mental images in mind. Such representative visual concepts are analogous to the notion of canonical viewpoints introduced by Palmer et al. [74], which is defined with the following four criteria: 1) which view do you like given a set of photos, 2) which view do you choose when taking a photo, 3) which view of the object is most recognizable and 4) when imagining an object, which view do you see. Their subsequent user experiments [74] demonstrated that people's preferences largely agree with each other, and they choose similar types of view regardless of the four questions asked. We propose to explore the notion of canonical visual concepts through users studies in the manner similar to Palmer's work [1]

Figure 6 illustrates a possible notion of a canonical visual concept. Among a set of images that are related to Mona Lisa, some of them, such as the near-duplicates of the original painting shown in the middle, may be perceived as "canonical" more than others. In fact, this illustration is automatically generated from the search results using the query "Mona Lisa." The distances among the thumbnails are inversely correlated with the similarities computed from the image features. The largest thumbnails represent the images most "central" with respect to the rest of the images in the collection. Less central images are shown in Figure 7.

This work conjectures that the *centrality* of a photo, relative to others produced by a Web retrieval system, is correlated with the likelihood of the photo being considered

[1] We started our investigation in the summer of 2006, introduced the notion of learning "canonical image" from Web search results [47] in 2007, and proposed a scalable approach to compute such scores [46] in 2008.

Figure 6: A two-dimensional projection of the search results produced by the query "Mona-Lisa," where the distances among the thumbnails are inversely correlated with the similarities computed from the image features. The largest thumbnails represent the images most "central" with respect to the rest of the images in the collection.

Figure 7: Less "central" images for "Mona-Lisa."

canonical or relevant by people. If this conjecture is true, then one can automatically derive centrality measurement from Web photos and use it to improve real-life image retrieval tasks. Chapter 5 will conduct experiments to evaluate whether selecting centrality images can improve user performance on specific retrieval tasks.

This chapter addresses two questions associated with computing the centrality score of the Web images. First, how can we reliably and *efficiently* estimate the centrality of images on the Web. Second, can we demonstrate that such centrality measurement of Web images is correlated with the notion of being canonical from the perspective of the users.

Section 2.3 presents a scalable approach to compute centrality scores from the Web images. In particular, we first compute the pairwise image similarity for search results based on efficient hashing [22] of SIFT [62] features, then use the principle Eigenvector of the adjacency matrix (which can be computed iteratively and in parallel using power iteration method) to measure the centrality of the images.

Section 2.5 presents user studies with centrality images derived from the Web. We conducted a large scale user studies with more than 1000 queries and 150 users, and demonstrated that images with high centrality scores are considered as more relevant to the query than those with low centrality scores.

This chapter is closely related with subsequent works described in later chapters. Chapter 3 presents a supervised learning approach to learn distance functions from user data, and Chapter 5 evaluates whether selecting images with high centrality scores can indeed improve user performance on specific image retrieval tasks. The contents of this chapter has been published in [47, 46].

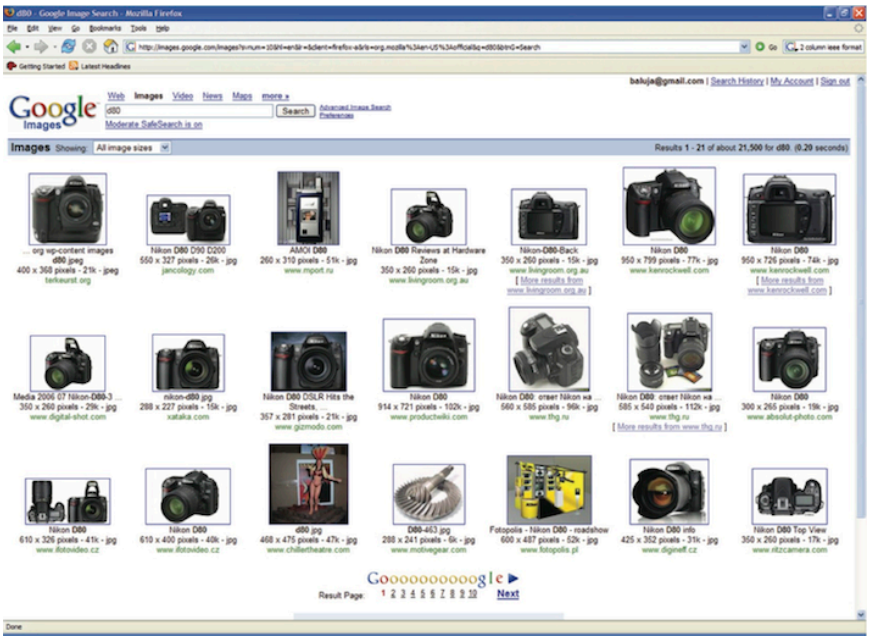
2.1 Introduction

Although image search has become a popular feature in many search engines, including Yahoo, MSN, Google, etc., the majority of image searches use very little, if any, image information. Due to the success of text-based search of web pages, and in part to the difficulty and expense of using image-based signals, most search engines return images solely based on the text of the pages from which the images are linked. For example, to

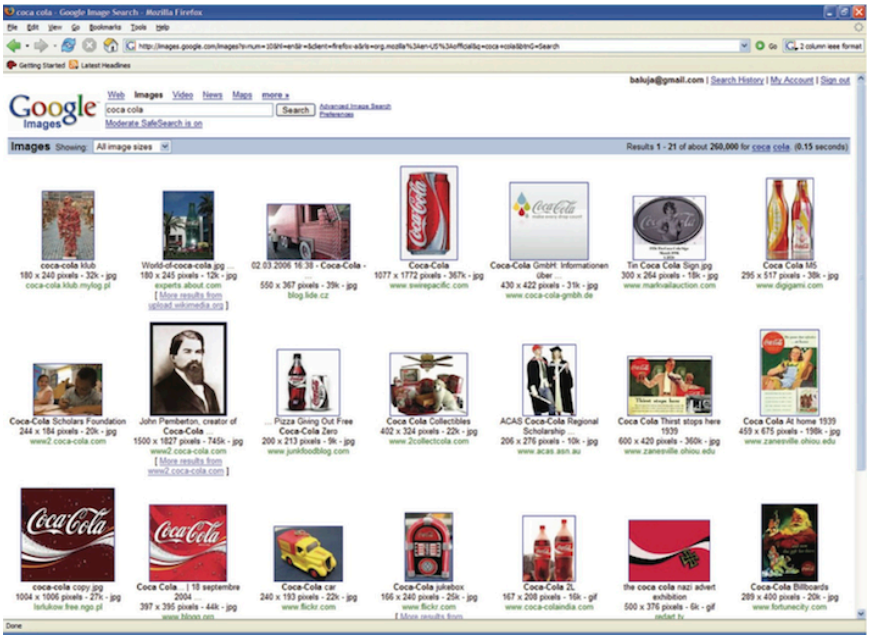
find pictures of the Eiffel Tower, rather than examining the visual content of the material, images that occur on pages that contain the term "Eiffel Tower" are returned. No image analysis takes place to determine relevance or quality. This can yield results of inconsistent quality. For example, the query "d80", a popular Nikon camera, returns good results as shown in Figure 8(a). However, the query for "Coca Cola" returns mixed results as shown in Figure 8(b) - the expected logo or Coca Cola can/bottle is not seen until the 4th result. This is due in part to the difficulty in associating images with keywords, and in part to the large variations in image quality and user perceived semantic content.

This work studies the hypothesis that images that are similar to other images in the search results are also considered as more relevant images to the query. The premise is simple: an author of a web page is likely to select images that, from his or her own perspective, are relevant to the topic. Rather than assuming that every user who has a web-page relevant to the query will link to an image that every other user finds relevant, our approach relies on the combined preferences of many web content creators. For example, in Figure 8(b), many of the images contain the familiar red Coca Cola logo. In some of the images, the logo is the main focus of the image, whereas in others it occupies only a small portion. Nonetheless, its repetition in a large fraction of the images returned is an important signal that can be used to infer a common "visual theme" throughout the set. Estimating the relative strength each image in representing the image collection, and study its relationship with user perceived relevance scores is the focus of this study.

It is not obvious that images that are similar to other images in the search results are perceived as relevant from the perspective of the users. It is also unclear whether ranking based on content-based features can be applied to hundreds of thousands of popular queries. For example, due to the high dimensionality of visual features and the difficulty in associating visual features with semantic content, it is not clear whether images with the most representative visual feature are considered as relevant and meaningful to the users. Also, recently proposed methods [27] that relies on probabilistic graphical models are expensive to train and sensitive to the choice of model parameters. For example, Web queries with multiple visual concepts such as "lincoln memorial" and "nemo" (shown in



(a) d80



(b) coca-cola

Figure 8: The query for "d80", a popular Nikon camera, returns good results on Google. However, the query for "Coca Cola" returns mixed results.

Figure 9: Many queries like "lincoln memorial" (first 2 images) and "nemo" (last 3 images) contain multiple visual themes.

Figure 9) can be particularly challenging for parts-based probabilistic models [27].

This work adopts a simple way to estimate how well individual images captures the overall visual content of the search results. We first compute the pairwise visual similarity among images using efficiently hashed SIFT [62] features, and use the principle Eigenvector of the adjacency matrix as a measurement of "centrality" that indicates how well each image represent the visual content of the search results.

Using pairwise image similarity as an intermediate representation of visual features has practical advantages for refining search results. First, it gives search engine designers the flexibility to customize image similarities through domain engineering. For example, similarity computations that capture higher order feature dependencies [2] and learning techniques can be efficiently employed [113, 108, 32, 94]. Further, even non-visual information, such as user-generated co-visitation [104, 5] statistics, can be easily combined with visual features to make similarity scores more semantically relevant [3]. Second, pairwise image similarities are easy to interpret, so the results can be visualized easily [4]. The simplicity and effectiveness of such approaches were demonstrated by He et al. [111], who first suggested combining PageRank with visual similarity for image retrieval, and was later extended by Hsu et al. [107] for video retrieval and Joshi et al. [54] in the development of "Story Picturing Engine."

[2] For example, geometric constraints [62] can be easily used in conjunction with local descriptors to reduce registration error.

[3] Chapter 3 presents methods to learn distance functions from user click-patterns.

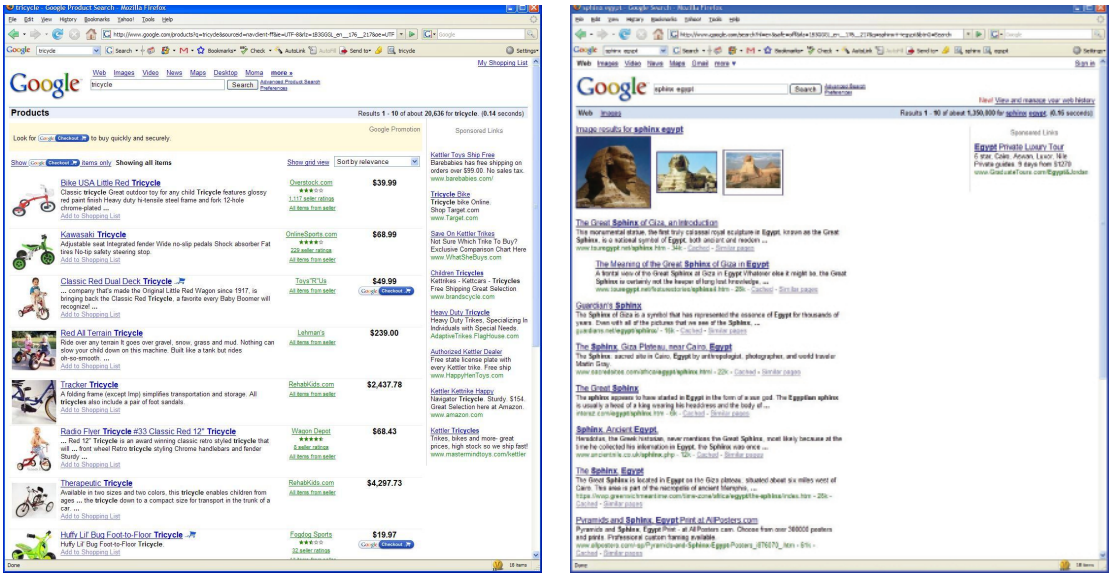
[4] Google Image Swirl [48] presents search results as a hierarchical exemplar tree computed from image similarities.

2.2 Related works

Our work belongs to the general category of Content-Based Image Retrieval (CBIR), an active research area driven in part by the explosive growth of personal photography and the popularity of search engines. A comprehensive survey on this subject can be found in [23]. Many systems proposed in the past [75, 96, 63, 9] are considered as "pure" CBIR systems – search queries are issued in the form of images and similarity measurements are computed exclusively from content-based signals. On the other hand, "composite" CBIR systems [54, 27] allow flexible query interfaces and a diverse set of signal sources, a characteristic suited for Web image retrieval as most images on the Web are surrounded by text, hyperlinks and other relevant metadata. For example, Fergus et al. [27] proposed the use of "visual filters" to re-rank Google image search results, bridging the gap between "pure" CBIR systems and text-based commercial search engines. These "visual filters" are learned from the top 1000 search results via parts-based probabilistic models [26], a form of Probabilistic Graphical Models (PGMs), to capture the higher order relationship among the visual features.

However, PGMs have several important limitations for Web image retrieval. First, as generative models are factored according to the structures of the model, a suboptimal model structure can significantly reduce modelling and especially the classification performance [31, 50]. An overly sparse model may neglect important higher order feature dependencies, while learning complex structures and their associated parameters are computationally prohibitive for large scale web image retrieval, and are prone to data noise, especially given the nature of the Web images and the diverse visual representation of object categories. Furthermore, there is an important mismatch between the goal of object category learning and image ranking. Object category learners are designed to model the relationship between features and images, whereas images search engines are designed to model the relationships (order) among images. Although a well trained object category filter can improve the relevancy of image search results, they offer limited capability to directly control how and why one visual theme, or image, is ranked higher than others.

Different from pure CBIR systems [96, 9, 63], VisualRank retains the commonly used



(a) Google product search (b) Mixed-Result-Type Search

Figure 10: In many uses, we need to select a very small set (1-3) of images to show from potentially millions of images. Unlike ranking, the goal is not to reorder the full set of images, but to select only the "best" ones to show.

text query interface and utilizes the visual similarities within the entire set of images for image selection. This approach complements pure CBIR systems in several ways: 1) text is still the most familiar, and often the only, query medium for commercial search engine users, 2) VisualRank can be effectively used in combination with other CBIR systems by generating a more relevant and diverse set of initial results, which often results in a better starting point for pure CBIR systems, 3) There are real-world usage scenarios beyond "traditional" image search where image queries are not feasible. In many uses, we need to select a very small set of images to show from potentially millions of images. Unlike ranking, the goal is not to reorder the full set of images, but to select only the "best" ones to show. Two concrete usage cases for this are: *1. Google product search*: only a single image is shown for each product returned in response to a product query; shown in Figure 10(a). *2. Mixed-Result Search*: to indicate that image results are available when a user performs a web (web-page) query, a small set of representative images may also be shown to entice the user to try the image search as shown in Figure 10(b). In both of these examples, it is paramount that the user is not shown irrelevant, off-topic, images. Finally, it is worth

noting that as good similarity functions are the foundation of CBIR systems, VisualRank can easily incorporate advances in other CBIR systems.

The recently proposed affinity propagation algorithm [30] also attempts to find the most representative vertices in a graph. Instead of identifying a collection of cluster centers, VisualRank differs from affinity propagation by explicitly computing the ranking score for all images. Several other studies have explored the use of a similarity based graph [58, 112] for semi-supervised learning. Given an adjacency matrix and a few labelled vertices, unlabelled nodes can be described as a function of the labelled nodes based on the graph manifolds. In this work, our goal is not classification; instead, we model the centrality of graph as a tool for ranking images. Another related work is by Zhu et al. [112], who proposes to use a random-walk model on graph manifolds to generate "smoothed" similarity scores that are useful in ranking the rest of the images when one of them is selected as query image. Our approach differs from [112] by generating an *a priori* ranking given a group of images.

Our work is closely related to [27], as both explore the use of content-based features to improve commercial image search engine. Random-walk based ranking algorithms were proposed by [111, 107, 54] for multimedia information retrieval; detailed comparison to these approaches were given in the previous section. The notion of selecting "canonical" images from the Web is related to [95, 56, 47] that computes image summarization from online photo collections. Simon et al. [95] proposes an unsupervised clustering technique to cluster images downloaded from Flickr, and use the cluster centroid as visual summaries.

2.3 System design

2.3.1 Visual Features

This work uses local descriptors to represent images. Comparing with global features such as color histograms and shape, local descriptors contain a richer set of image information and are relatively stable under different transformations and, to some degree, lighting variations. Such allows for more robust representation of regions in the images we are interested in. For example, as shown in Figure 11, the search results for "Golden Gate" often contain images taken from different locations, with different cameras, focal lengths, compositions,

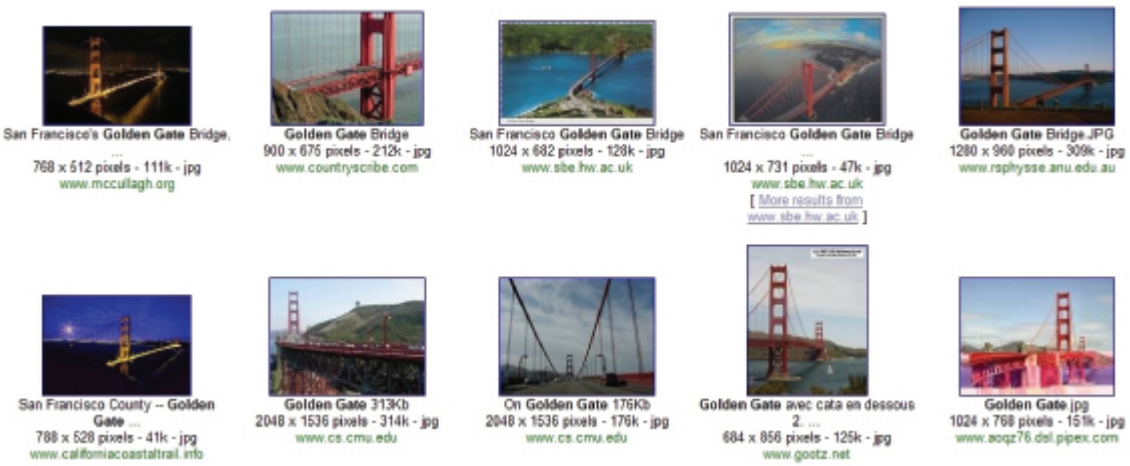


Figure 11: Similarity measurement must handle potential rotation, scale and perspective transformations.

etc. Examples of local features include Harris corners, Scale Invariant Feature Transform (SIFT) [62], Shape Context [6] and Spin Images [60] to name a few. Mikolajczyk and Schmid [65] presented a comparative study of various descriptors, [109] presented work on improving the their performance and computational efficiency. In this work, we use the SIFT features, with a Difference of Gaussian (DoG) interest point detector and orientation histogram feature representation as image features. However, any of the local features could have been substituted.

We used a standard implementation of SIFT; for completeness, we give the specifics of our usage here. A DoG interest point detector builds a pyramid of scaled images by iteratively applying Gaussian filters to the original image. Adjacent Gaussian images are subtracted to create Difference of Gaussian images, from which the characteristic scale associated with each of the interest points can be estimated by finding the local extrema over the scale space. Given the DoG image pyramid, interest points located at the local extrema of 2D image space and scale space are selected. A gradient map is computed for the region around the interest point and then divided into a collection of subregions from which an orientation histogram can be computed. The final descriptor is a 128 dimensional vector by concatenating 4x4 orientation histogram with 8 bins. Given two images, we define their similarity as the number of local features shared between them, divided by their. average

number of interest points.

2.3.2 Similarity computation through Locality Sensitive Hashing (LSH)

Instead of training an object category model [27] based on the signatures of the images, this works adopts a simpler approach that is to compute the pairwise similarity of the images, then compute a ranking score based on the distance between an image to its neighbors. Using pairwise image similarity allows one to abstract away from the features, and allow one to develop robust similarity function depending on the task at hand. However, the drawbacks of similarity is that it scales quadratically to the number of local features in the database. As we are computing ranking scores for each query, and each query usually retrieves less than 2000 images, it is possible to exhaustively compute pairwise similarity for images. However, to evaluate large number of queries, a more efficient way of matching local features is required.

A more efficient approach is to use a hash table to store all the local descriptors such that similar descriptors fall into the same bin [5]. In the extreme case, where only exact duplicate are considered as matches, one can simply use the original descriptor value as hash key (by converting the 128 dimensional vector into a single long hash key). To match "similar", non-exact-duplicate, local descriptors under different lighting conditions and other variations, a more relaxed, distance preserving, hashing function can be used.

Matching local descriptors efficiently has received tremendous research attention in the recent years [71, 114, 55, 90, 72, 76]. In particular, Nister et al. [71] proposed the use of "visual vocabularies," a set of distinctive quantitized local descriptors learned via hierarchical k-mean clustering. Raw features are mapped into visual vocabularies by traversing down vocabulary tree to find the closest leaf. This process can be viewed as constructing a hash function to map raw descriptors into a key, in this case, the visual vocabulary.

For our algorithm, approximation methods [22, 44, 76] to measure similarity are sufficient. Because VisualRank relies on the global structure of the graph to derive its ranking,

[5] Due to the memory requirement, hashing is practical only for a limited number of local features.

we expect it to be robust against localized noise (mismatch or missed matches of local features in images). Intuitively, if the distance measurement captures an overall notion of user perceived similarity, small difference in the magnitudes of the distance will have negligible effect on the end results. We will use a version of the Locality-Sensitive Hashing (LSH) approximate matching approach.

LSH is an approximate kNN technique introduced by Indyk and Motwani [44]. LSH addresses the similarity match problem, termed $(r;\epsilon)$-NN, in sub-linear time. The goal, stated formally, is as follows: given a point q (query) in d-dimensional feature space, for exact kNN: for any point q, return the point p that minimizes $D(p;q)$. For approximate kNN: if there exists an indexed point p such that $D(p;q) \leq r$, then with high probability return an indexed point that is of distance at most $(1+\epsilon)r$. If no indexed point lies within $(1+\epsilon)r$ of q, then LSH should return nothing, with high probability. Sukthankar et al. [114] have explored LSH in the task of near-duplicate image detection and retrieval and obtained promising results. The particular hash function in [114] was best suited for the preservation of Hamming distance; for our work, we follow the the recent work of Datar et al. [22]. [22] has proposed hash function for l_2 norms, based on p-stable distributions [43]. Here, each hash function is defined as:

$$h_{a,b}(V) = \lfloor \frac{aV+b}{W} \rfloor \tag{1}$$

where a is a d-dimensional random vector with entries chosen independently from a Gaussian distribution and b is a real number chosen uniformly from the range [0, W]. W defines the quantization of the features, and V is the original feature vector. Equation 1 is very simple to implement and efficient.

In practice, best results are achieved by using L number of hash tables rather than a single one. For each hash tables, we reduce the collision probability of non-similar objects by concatenating K hash functions. Two features are considered as a match if they were hashed into the same bin in C out of the L hash tables; effectively, this provides a means of setting a minimum match threshold, thereby eliminating coincidental matches that occur in only a few of the tables. We group all the matched features by their associate image, and the similarity matrix, S is computed by the total number of matches normalized by their

average number of local features. The exact parameter settings are given below.

2.3.3 Computing VisualRank scores

Given a graph with vertices and a set of weighted edges, one way to measure how well each vertex represent the graph is to compute its centrality scores. The cardinality of the vertice or the sum of geodesic distance to the surrounding nodes are all variations of centrality measurement. In this work, we use Eigenvector centrality to represent the visual content of the images.

As an example of a successful application of Eigenvector Centrality, PageRank [8] pre-computes a rank vector to estimate the importance for all of the webpages on the Web by analyzing the hyperlinks connecting web documents [6]. Intuitively, pages on Amazon.com are important with many pages pointing to them. Pages pointed to by Amazon.com may therefore, also have high importance. Non-uniform damping vectors were suggested previously by Haveliwala [38] to compute topic-biased PageRank for web documents.

Eigenvector Centrality is defined as the principle Eigenvector of a square stochastic adjacency matrix, constructed from the weights of the edges in the graph. It has an intuitive Random Walk explanation: the ranking scores correspond to the likelihood of arriving in each of the vertices by traversing through the graph (with a random starting point), where the decision to take a particular path is defined by the weighted edges.

In this work, in order to highlight the application to visual features, we refer the eigenvector as **visual-rank** of the images. VisualRank employs the Random Walk intuition to rank images based on the visual-hyperlinks among the images. The intuition of using these visual-hyperlinks is that if a user is viewing an image, other related (similar) images may also be of interest. In particular, if image u has a visual-hyperlink to image v, then there is some probability that the user will jump from u to v. Intuitively, images related to the query will have many other images pointing to them, and will therefore be visited often (as long as they are not an isolated and in a small clique). The images which are visited often

[6] The PageRank vector can be pre-computed and be independent of the search query. Then, at query time, PageRank scores can be combined with query-specific retrieval scores to rank the query results. This provides a faster retrieval speed than many query-time methods [57].

are deemed important. Further, if we find that an image, v, is important and it links to an image w, it is casting its vote for w's importance – because v is itself important, the vote should count more than a "non-important" vote.

VisualRank (VR) is iteratively defined as the following:

$$VR = S^* \times VR \tag{2}$$

S^* is the column normalized, symmetrical adjacency matrix S where $S_{u,v}$ measures the visual similarity between image u and v. Since we assume similarities are commutative, the similarity matrix S is undirected. Repeatedly multiplying VR by S^* yields the dominant eigenvector of the matrix S^*. Although VR has a fixed point solution, in practice it can often be estimated more efficiently through iterative approaches.

VisualRank converges only when matrix S^* is aperiodic and irreducible. The former is generally true for the web, and the later usually requires a strongly connected graph, a property guaranteed in practice by introducing a damping factor d into Equation 2. Given n images, VR is defined as:

$$VR = dS^* \times VR + (1-d)p, \quad \text{where } p = [\frac{1}{n}]_{n\times 1}. \tag{3}$$

This is analogous to adding a complete set of weighted outgoing edges for all the vertices. Intuitively, this creates a small probability for a random walk to go to some other images in the graph, although it may not have been initially linked to the current image. $d > 0.8$ is often chosen for practice; empirically, we have found the setting of d to have relatively minor impact on the global ordering of the images.

In place of the uniform damping vector p in Equation 3, we can use a non-uniform vector q to bias the computation. For example, we can use it to increase the effect of images ranked high in the initial search engine results, since they are selected, albeit through non-visual features, to be the best match to the query. Vector q can be derived from image quality, anchor page quality, or simply the initial rank from commercial search engines. The intuition is that "random surfers" are more likely to visit and traverse through images that have higher prior expectation of being relevant. For example, if we assume the top m search

results from commercial search engines to be of reasonable quality, we can use $q = v_j$, where

$$v_j = \begin{cases} \frac{1}{m}, & j \leq m \\ 0, & otherwise \end{cases} \tag{4}$$

As an example of a successful application of Eigenvector Centrality, PageRank [8] pre-computes a rank vector to estimate the importance for all of the webpages on the Web by analyzing the hyperlinks connecting web documents [7]. Intuitively, pages on Amazon.com are important with many pages pointing to them. Pages pointed to by Amazon.com may therefore, also have high importance. Non-uniform damping vectors were suggested previously by Haveliwala [38] to compute topic-biased PageRank for web documents.

2.3.4 Query Dependent Visual Rank

It is computationally infeasible to generate the similarity graph S for the billions of images that are indexed by commercial search engines. One method to reduce the computational cost is to precluster web images based using metadata such as text, anchor text, similarity or connectivity of the web pages on which they were found. For example, images associated with "Paris", "Eiffel Tower", "Arc de Triomphe" are more likely to share similar visual features than random images. To make the similarity computations more tractable, a different VisualRank can be computed for each group of such images.

A practical method to obtain the initial set of candidates mentioned in the previous paragraph is to rely on the existing commercial search engine for the initial grouping of semantically similar images. For example, similar to [27], given the query "Eiffel Tower" we can extract the *top-N* results returned, create the graph of visual similarity on the N images, and compute VisualRank only on this subset. In this instantiation, VisualRank is query dependent; although the VisualRank of images in the N images is indicative of their importance for answering the query, the same image may have a different score when it is a member of a different set of images that is returned in response to a different query. In the experiment section, we follow this procedure on 2000 of the most popular queries for

[7] The PageRank vector can be pre-computed and be independent of the search query. Then, at query time, PageRank scores can be combined with query-specific retrieval scores to rank the query results. This provides a faster retrieval speed than many query-time methods [57].

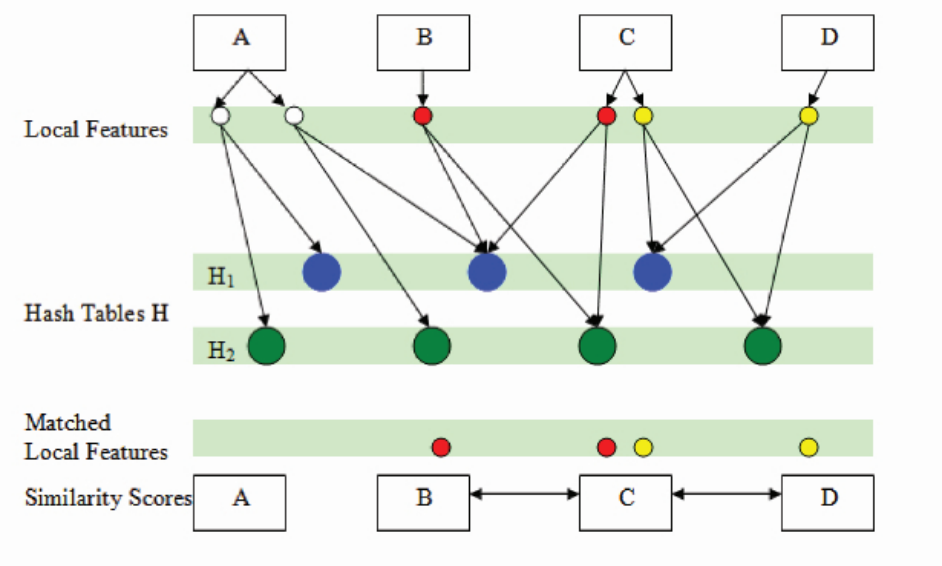


Figure 12: A visual representation of our hashing-scheme. Local features extracted from a collection of images are hashed into a collection of LSH hash tables. Features hashed into the same bin in multiple hash families are considered matches, and contribute to the similarity score between their corresponding images.

Google Product Search.

2.3.5 Summary of the system

The VisualRank system can be summarized as the following 4 steps. A visual representation of the process is given in Figure 12.

1. Local features are generated for a group of images, scaled to have a maximum axis size of 500 pixels. From our study, 1000 web images usually contain 300,000 to 700,000 feature vectors.

2. A collection of L hash tables $H = H_1, H_2, ..., H_L$ are constructed, each with K number of hash functions as shown in Equation 1. Each of the descriptors is indexed into each of the hash-tables. Empirically, we determined that $L = 40$, $W = 100$, and $K = 3$ give good results.

3. For each descriptor, we aggregate objects with identical hash keys across L hash tables. Descriptors that share the same key in more than C hash-tables are considered as a match (C=3).

4. We regroup matched features by the images they are associated with. Optionally, for two images and their associated matching feature points, we use a Hough Transform to enforce a loose geometric consistency. A 4 dimensional histogram is used to store the "votes" the pose space (translation, scaling and rotation). At the end, we select the histogram entry with the most votes as the most consistent interpretation. The surviving matching points are used to compute the similarity score.

5. A pair of images are considered as a match if they share more than 3 matched descriptors. The similarity of two images is computed by the total number of matches normalized by their average number of local features.

6. Given similarity matrix S, we can use VisualRank algorithm to generate the top N images.

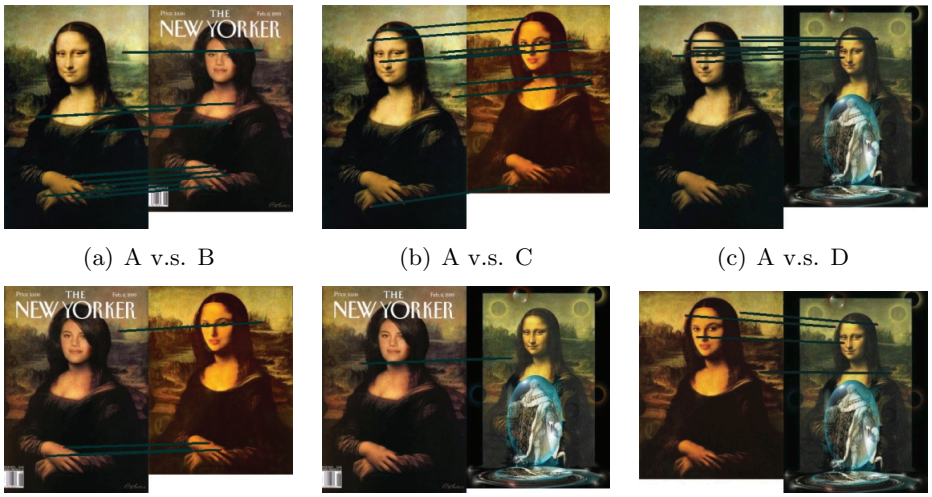
With the techniques mentioned above, and non-optimized code, it takes approximately 10 minutes to compute and hash the local descriptors for 1000 images, and an additional 5 minutes is required to compute the full similarity matrix. Although this is a significant computational requirement, it allows us to pre-compute the results to many popular queries. For example, with 1000 modest CPUs, the VisualRank for the top 100,000 queries can be computed in less than 30 hours.

2.4 Case studies

This section selected a few queries to illustrate, through the visualization of search results, how visual coherency can be used to improve retrieval relevancy.

2.4.1 Queries with homogeneous visual concepts

VisualRank improves the relevance of image search results under queries with homogeneous visual concepts. This is achieved by identifying the vertices that are located at the "center" of weighted similarity graph. "Mona-lisa" is a good example of a search query with a single homogeneous visual concept. Although there are many comical variations (i.e. "Bikini-lisa", "Monica-Lisa"), they are all based on the original painting. As shown in Figure 13, the original painting contains more matched local features than others, thus has the highest



(a) A v.s. B (b) A v.s. C (c) A v.s. D

(d) B v.s. C (e) B v.s. D (f) C v.s. D

Figure 13: Since all the variations (B, C, D) are based on the original painting (A), A contains more matched local features than others.

likelihood of visit by an user following these probabilistic visual-hyperlinks. Figure 14 is generated from the top 1000 search results of "Mona-Lisa." The graph is very densely connected, but not surprisingly, the center of the images all correspond to the original version of the painting.

2.4.2 Queries with heterogeneous visual concepts

VisualRank can improve the relevancy and diversity of queries that contain multiple visual concepts. Examples of such queries that are often given in information retrieval literature include "Jaguar" (car and animal) and "Apple" (computer and fruit). However, when considering images, many more queries also have multiple canonical answers. For example, the query "Lincoln Memorial", shown in Figure 15, has multiple good answers (pictures of the Lincoln Statue, pictures of the building, etc). In practice, VisualRank is able to identify a relevant and diverse set of images as top ranking results; there is no *a priori* bias towards a fixed number of concepts or clusters.

An interesting question that arises is whether simple heuristics could have been employed for analyzing the graph, rather than using a VisualRank / Eigenvector approach. For example, a simple alternative is to select the high degree nodes in the graph, as this implicitly captures the notion of well-connected images. However, this fails to identify the different

Figure 14: Similarity graph generated from the top 1000 search results of "Mona-Lisa." The largest two images contain the highest VisualRank.

Figure 15: Top ten images selected by VisualRank from the 1000 search results of "Lincoln Memorial." By analyzing the link structure in the graph, VisualRank identifies a highly relevant yet diverse set of images. (A) Night time photo of the Lincoln Statue. (B) Daytime photo of the statue. (C) Lincoln Memorial Building.

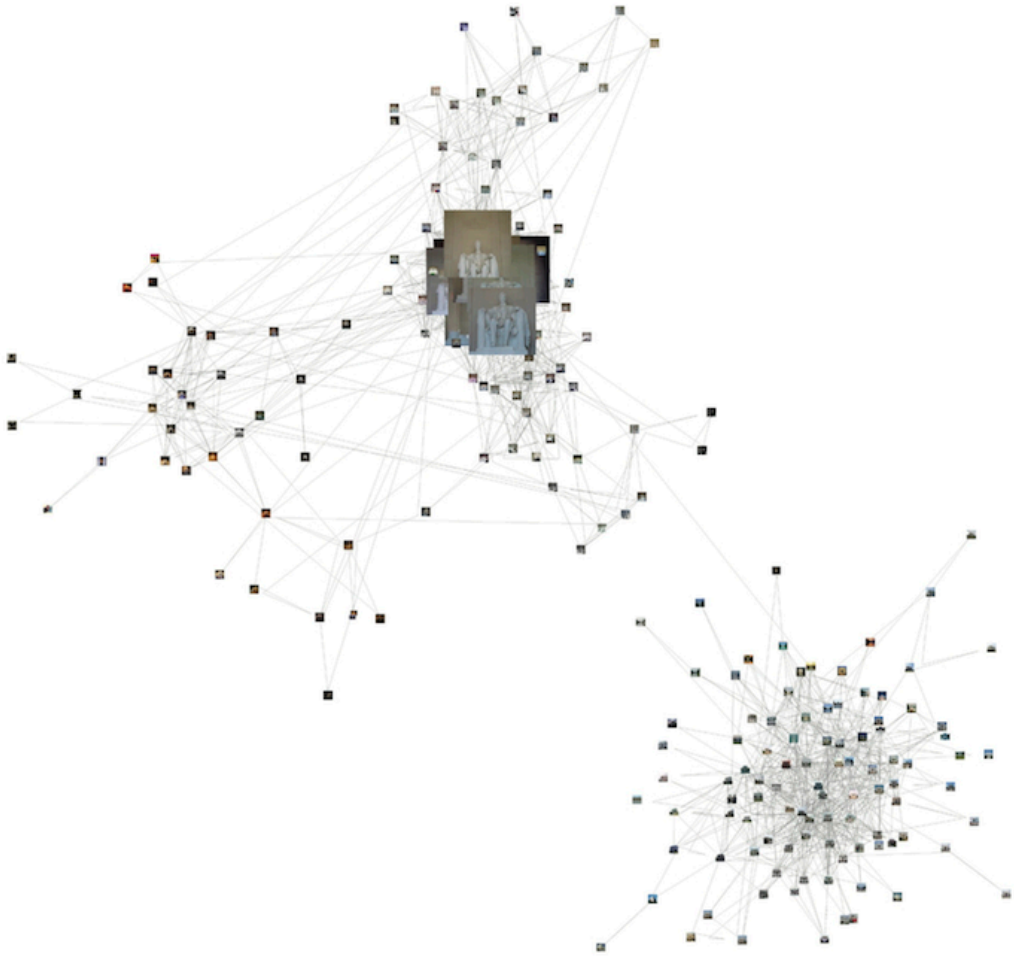
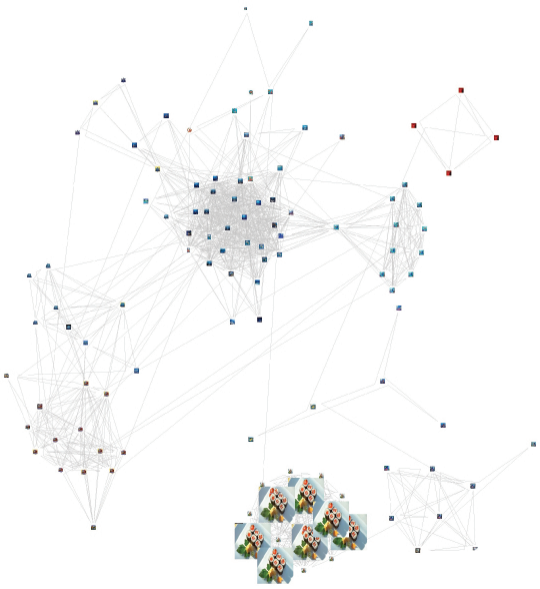


Figure 16: Alternative method of selecting images with the most "neighbors" tend to generate relevant but homogeneous set of images.

(a) VisualRank



(b) High Degree images

(c) Spam image

Figure 17: By analyzing the global structure of the graph, VisualRank (a) avoids selecting images simply because they are close-duplicates of each other. The alternative methods of selecting images with the highest weighted degree is susceptible to this (b), as it find the spam images (c) repeatedly.

distinctive visual concepts as shown in Figure 16. Since there are more close matches of "Lincoln statue," they reinforce each other to form a strongly connected clique. Further, the random-walk model also accounts for distracting or "spam" images, as will be shown in the next section. Of course, measures can be added to detect these cases; however, VisualRank provides a principled and intuitive method, through a simple fixed point computation, to capture these insights.

2.4.3 Performance in the presence of distracting images

Visual similarity among images offers tremendous information about the popularity and relevance of a particular image. However, it can be susceptible to manipulations. For example, in commercially deployed systems, adversary content creators can inflate the rankings of their own images by placing a large quantity of duplicate images on the web [8]. Although those images may bring additional traffic to their website, users may not find them helpful. This practice is analogous to "Link Spam" on the web, where artificially constructed densely connected webpage are used to inflate their rankings in regular search engines.

Even in its straightforward implementation, VisualRank is resistant to many forms of similarity link-spam by analyzing the global structure of the graph. For example, the top 1000 images collected with query "nemo" contained many (near/exact)-duplicated images of "Nemo Sushi" shown in Figure 17(c). Note that these images reinforce each other. Simpler algorithms, such as selecting high degree nodes, are easily misled, as shown in Figure 17(b). VisualRank is resistant to this due to the normalization of similarity matrix in Equation 2; first, there is not a strong probability of visitation in a random-walk model, unless the visitor is already on one of the images (the set of images is not well connected to the rest of the graph), second, although these images form a tight clique, they have an equal share of transitional probability, thus there is no "authority" node within the set. Other more distinctive images with a more diverse set of matches are selected with VisualRank as shown in Figure 17(a).

[8] Note that "adversary" is meant literally; it is common practice for content creators to submit many duplicate or near-duplicate images, web pages, etc. intentionally designed to bias ranking algorithms to place their content above others.

2.5 Experiments

To ensure that our algorithm works in practice, we conducted experiments with images collected directly from the web. In order to ensure that the results would make a significant impact in practice, we concentrated on the 1000 most popular product queries [9] on Google (product search). Typical queries included "ipod", "xbox", "Picasso", "Fabreze", etc [10]. For each query, we extracted the top 1000 search results from Google image search in July, 2007, with the strict safe search filter. The similarity matrix is constructed by counting the number of matched local features for each pair of images after geometric validation normalized by the number of descriptors generated from each pairs of images.

It is challenging to quantify the quality of (or difference of performance) of sets of image search results for several reasons. First, and foremost, user preference to an image is heavily influenced by a user's personal tastes and biases. Second, asking the user to compare the quality of a *set* of images is a difficult, and often a time consuming, task. For example, an evaluator may have trouble choosing between group A, containing five relevant but mediocre images, and group B, that is mixed with both great and bad results. Finally, assessing the differences in ranking (when many of the images between two rankings being compared are the same) is error-prone and imprecise, at best. Perhaps the most principled way to approach this task is to build a global ranking based on pairwise comparisons. However, this process requires significant amount of user input, and is not feasible for large numbers of queries.

To accurately study the performance of VisualRank, subject to practical constraints, we devised two evaluation strategies. Together, they offer a comprehensive comparison of two ranking algorithms, especially with respect to how the rankings will be used in practice.

[9] The most often queried keywords during a period in August.

[10] We chose product (and travel/landmark) related queries for three reasons. First, they are extremely popular in actual usage. Second, they lend themselves well to the type of local feature detectors that we selected in this study (in Section 7 we describe other categories of queries that may benefit from alternative sets of image features). Third, users have strong expectations of what results we should return for these queries; therefore, this provides an important set of examples that we need to address carefully.

2.5.1 Experiment I: User study on retrieval relevancy

This study is designed to study a conservative version of "relevancy" of our ranking results. For this experiment, we mixed the top 10 VisualRank selected images with the top 10 image from Google, removed the duplicates, and presented them to the user. We asked the user: "Which of the image(s) are the least relevant to the query?" [11] For this experiment, more than 150 volunteer participants were chosen, and were asked this question on a set of randomly chosen 50 queries selected from the top-query set. There was no requirement on the number of images that they marked.

There are several interesting points to note about this study. First, it does not ask the user to simply mark relevant images; the reason for this is that we wanted to avoid a heavy bias to a user's own personal expectation (i.e. when querying "Apple" did they want the fruit or the computer?). Second, we did not ask the users to compare two sets; since, as mentioned earlier, this is an arduous task. Instead, the user was asked to examine each image individually. Third, the user was given no indication of ranking; thereby alleviating the burden of analyzing image ordering.

In order to quantify the effectiveness of visual features, VisualRank was computed with a uniform bias vector, ignoring order of placement in the original search results. We measured the results for Google and VisualRank for three settings: the number of *irrelevant* images in the top-10, top-5, and top-3 images returned by each of the algorithms. Table 1 contains the comparison results. Among the top 10 images, VisualRank produced an average of 0.47 irrelevant results, this is compared with 2.82 by Google; this represents an 83% drop in irrelevant images. When looking at the top-3 images, the number of irrelevant images for VisualRank dropped to 0.20, while Google dropped to 0.81.

In terms of overall performance on queries, as shown in Table 2, VisualRank contains less irrelevant images than Google for 762 queries. In only 70 queries did VisualRank produce worse results than Google. In the remaining 168 queries, VisualRank and Google tied (in

[11] Typically given a query image or a text query, a set of ground-truth images are selected as either relevant or irrelevant with the following guideline (TREC): "if you were writing a report on the subject of the topic and would use the information contained in the document in the report, then the document is relevant."

Table 1: Relevancy Study

"Irrelevant" images per product query	**VisualRank**	**Google Images(2006)**
Among top 10 results	0.47	2.82
Among top 5 results	0.30	1.31
Among top 3 results	0.20	0.81

Table 2: Relevance comparison per query

	VisualRank	**Google**
Outperforming product queries	762	70

the majority of these, there were no irrelevant images). Figure 18 provides a query-by-query analysis between VisualRank and existing Google image search. The Y axis contains the number of "irrelevant" images, and the X axis lists the type of queries. The order of queries are sorted by number of "irrelevant" images retrieved by Google image search engine for better visualization.

To present a complete analysis of VisualRank, we describe two cases where VisualRank did not perform as expected. VisualRank sometimes fails to retrieve relevant images as shown in Figure 19. The first three images are the logos of the company which manufactured the product being searched for. Although the logo is somewhat related to the query, the evaluators did not regard them as relevant to the specific product for which they were searching. The inflated logo score occurs for two reasons. First, many product images contains the company logos; either within the product itself or in addition to the product. In fact, extra care is often given to make sure that the logos are clearly visible, prominent, and uniform in appearance. Second, logos often contain distinctive patterns that provides a rich set of local descriptors that are particularly well suited to SIFT-like feature extraction.

A second, but less common, failure case is when screen-shots of web pages are saved as images. Many of these images include browser panels or Microsoft Window's control panels that are consistent across many images. It is suspected that these mismatches can easily be filtered by combining VisualRank with other source of quality scores or measuring distinctiveness of the features not only within queries but also across queries; in a manner similar to using TF-IDF [87] weighting in textual relevancy. In fact, as shown in the next sections, some of the mismatches can be easily filtered by biasing the computation of

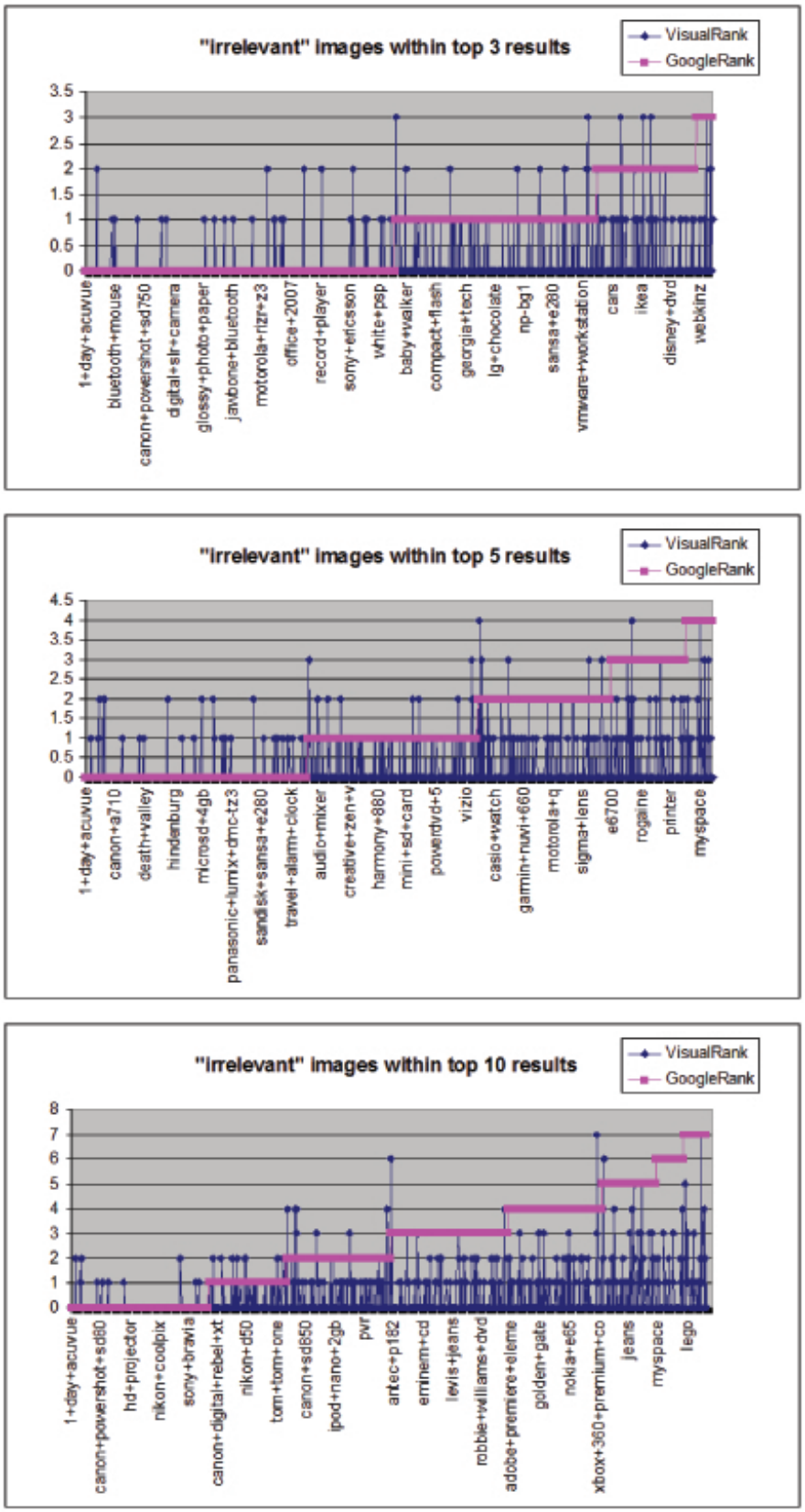


Figure 18: Number of irrelevant images (per query) retrieved by competing algorithms. For visual clarity, a sub-sample of corresponding queries (cars, etc) are shown under the x-axis. The queries are sorted by the number of irrelevant results retrieved by Google image search.

(a) dell computer (b) nintendo wii system (c) 8800 Ultra

(d) keychain (e) ps2 network adapter (f) dell computer

Figure 19: The particular local descriptors used provided a bias to the types of patterns found. These VisualRank selected images received the most "irrelevant" votes from the users for the queries shown.

Table 3: Relevancy Study

"Irrelevant" images per product query	**VisualRank$_{bias}$**	**VisualRank**	**HeuristicRank**
Among top 20 results	0.23	0.83	1.93
Among top 10 results	0.17	0.47	1.42
Among top 5 results	0.12	0.30	0.86
Among top 3 results	0.04	0.20	0.65

VisualRank with the initial order of placement from Google image search results.

2.5.2 Experiment II: Satisfaction and Click Measurement

Results from Experiment I show that VisualRank can effectively decrease the number of irrelevant images in the search results. However, user satisfaction is not purely a function of relevance; for example, numerous other factors such as diversity of the selected images must also be considered. Assuming the users usually click on the images they are interested in, an effective way to measure search quality is to analyze the total number of "clicks" each image receives.

We collected clicks for the top 40 images (first two pages) presented by the Google search results on 130 common product queries. The VisualRank for the top-1000 images for each of the 130 queries is computed and the top-40 images are reranked using VisualRank. To determine if the ranking would improve performance, we examine the number of clicks each method received from only the top-20 images (these are the images that would be displayed in the first page of results (on http://images.google.com)). The hope is that by reordering the top-40 results, the best images will move to the top; and would be displayed on the first page of results. If we are successful, then the number of clicks for the top-20 results under reordering will exceed the number of clicks for the top-20 under the default ordering.

It is important to note that this evaluation contains an *extremely severe bias that favors the default ordering*. The groundtruth of clicks an image receives is a function not only of the relevance to a query and quality of the image, *but of the position in which it is displayed*. For example, it is often the case that a mediocre image from the top of the first page will receive more clicks than a high quality image from the second page (default ranking 21-40). If VisualRank outperforms the existing Google Image search in this experiment, we can

Table 4: Relevancy Study

"Irrelevant" images per landmark query	VisualRank	Google Images(2006)
Among top 10 results	0.35	3.64
Among top 5 results	0.18	1.73
Among top 3 results	0.03	0.94

Table 5: Relevance comparison per query

	VisualRank	Google
Outperforming landmark queries	46	2

expect a much greater improvement in deployment.

When examined over the set of 130 product queries, the images selected by VisualRank to be in the top-20 would have received approximately 17.5% more clicks than those in the default ranking. This improvement was achieved despite the positional bias that strongly favored the default rankings.

2.5.3 An Alternate Query Set: Landmarks

To this point, we have examined the performance of VisualRank on queries related to products. It is also interesting to examine the performance on an alternate query set. Here, we present the results of an analogous study to the product-based one presented to this point; this study is conducted with common landmark related queries.

For this study, we gathered 80 common landmark related queries. Typical queries included: "Eiffel Tower", "Big Ben", "Coliseum" and "Lincoln Memorial". Similarly to product queries, these queries have rigid, canonical objects that are central to the answer. Table 4 shows the performance of VisualRank when minimizing the number of irrelevant queries in the top-10, top-5 and top-3 results. As was seen in the experiments with product images, VisualRank significantly outperforms the default rankings at all of the measured settings. Table 5 shows the number of queries that VisualRank outperformed Google and vice-versa. Note that the default Google rankings rarely outperformed VisualRank; however, there were a large number of ties (32), in which Google and VisualRank had an equal number of irrelevant images.

For the last measurement, we examine the clicks that would have been received under

VisualRank based reordering and under default settings. In 50 of the queries, VisualRank would have received more clicks, while in 27 of the queries the default ranking would have. The remaining 3 queries tied.

2.6 Conclusions

This work presents VisualRank, a scalable approach to compute image centrality scores for web search results, and demonstrate that image centrality scores are highly correlated with what users consider to be relevant to the query. The result was an approach that was able to outperform the default Google ranking on the vast majority of queries tried while maintaining reasonable computational efficiency for large-scale deployment. Importantly, the ability to reduce the number of irrelevant images shown is extremely important not only for the task of image ranking for image retrieval applications, but also for applications in which only a tiny set of images must be selected from a very large set of candidates.

CHAPTER III

LEARNING QUERY-SPECIFIC DISTANCE FUNCTION FROM CLICK PATTERNS

3.1 Introduction

Estimating image distances is central to all content-based image retrieval systems. For example, in Chapter 2, the centrality measurement is computed by first measuring pairwise distances of the images in the search results. Commonly used distance functions for image retrieval include Euclidean distance and Earth Mover distance [85]. In some cases, distance functions are learned from data [91, 113, 110]. Such methods have been generally adopted to learn a single distance for all images in the training data.

This work studies the problem of learning distance functions to be used in a hybrid image retrieval systems such as the one used by Google or Bing Images. Shown in Figure 20, such systems adopt a hybrid search approach in which a text-based query (e.g. "Eiffel Tower") is used to retrieve a set of relevant images, which are then refined by the user (e.g. by re-ranking the retrieved images based on similarity to a selected example). Unlike standard content-based image retrieval system, the goal is not to retrieve an similar image from the Web, but rather to re-rank the search results once an initial set of images are retrieved with

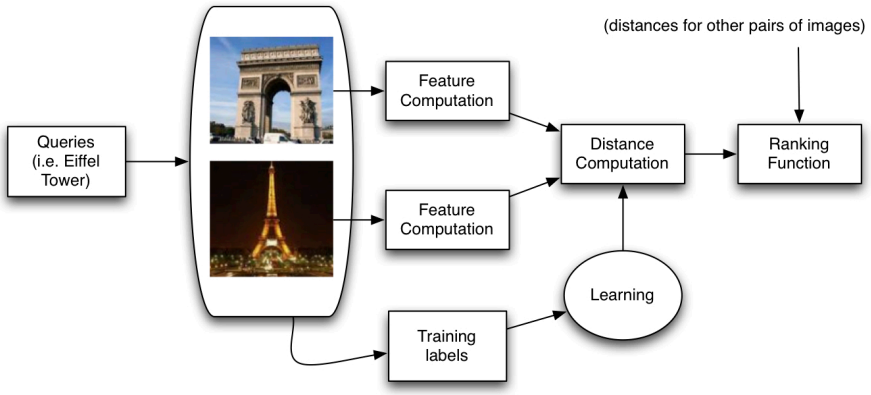
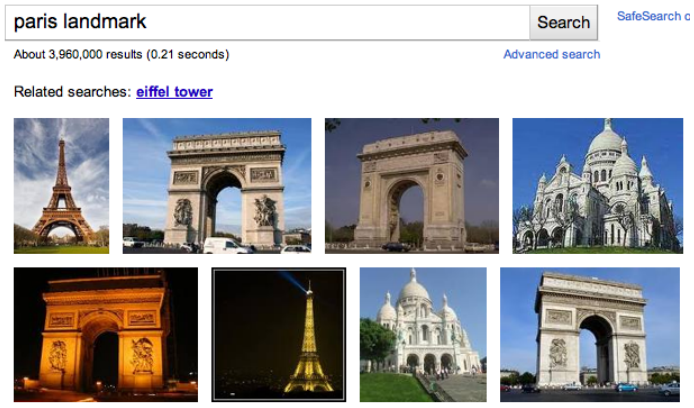
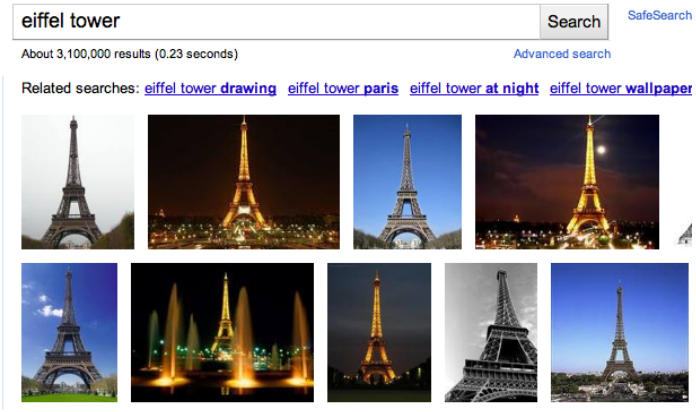


Figure 20: Role of distance functions in hybrid image retrieval systems



(a) Paris landmark



(b) Eiffel Tower

Figure 21: Top search results with the query *Paris landmarks* and *Eiffel Tower*

a text query. Therefore, we are interested in learning *query-specific* distance functions.

The motivation for learning query-specific distance functions stems from the fact that the appropriate choice of feature depends upon the query. For example, consider the problem of identifying a photo of Eiffel Tower. If the query is "Paris landmarks" as shown in Figure 21(a), then shape feature will be valuable as it differentiate Eiffel tower more clearly from other architectural structures. On the other hand, if the query is "Eiffel Tower" as shown in Figure 21(b), then color feature would be relatively more useful than shape. Since the context (e.g. the query "Eiffel Tower") is expected to already restrict the images to the correct landmark, the measure of similarity should instead group the images on a less constrained dimension, such as time-of-day, as the color distribution corresponding to time of day.

This work proposes to learn query-specific distance functions by adopting the large-margin learning approach introduced by [91]. We conjecture for the task of comparing images (retrieved by using text-based search engine), the relative importance of various image features depends on the particular query used. Therefore, instead of learning a single set of feature weights for all images (global distance function), we propose to learn separate feature weights for *each query*. This work is also closely related to the work by Frome et al. [33], where a separate distance function is learned for each query image. Learning distance per image, however, is not feasible for Web image retrieval as it requires collecting sufficient training data and caching the learned weights for *each* potential query image. From the observation that the frequency of search queries term usually follows power law distribution [89], our proposed query-specific approach needs only to be applied to images retrieved by the most popular queries and still service significant proportion of the search engine traffic.

The main challenge of learning image similarity for each query is to collect sufficient amount of training data. Standard ways to collect training information, such as manual image labeling of images [35, 25, 2] or relevance feedback [119, 67, 21] are costly as it requires active human participation. Also, manually assigned image labels are often not sufficiently descriptive of the images [79]. Instead of collecting training data based on explicit user participation, this work proposes to observe search engine users' click-patterns made during the process of conducting Web image search, which are captured by the search engine query logs.

A key distinction of our proposed approach is the use of the query logs of one type of image retrieval system (text-query based search engine) as training data for another type of image retrieval system (hybrid search engine), as shown in Figure 22. This approach allows one to leverage large quantities of feedback data from popular incumbent Web retrieval system, and use it to improve the performance of a new system that may not have sufficient relevance feedback data of its own. However, the challenge of using the logs of text-based Web search is that users are making image selections in an *unconstrained* and *open* system, and without a predefined guideline on what task to accomplish. This section is to study the

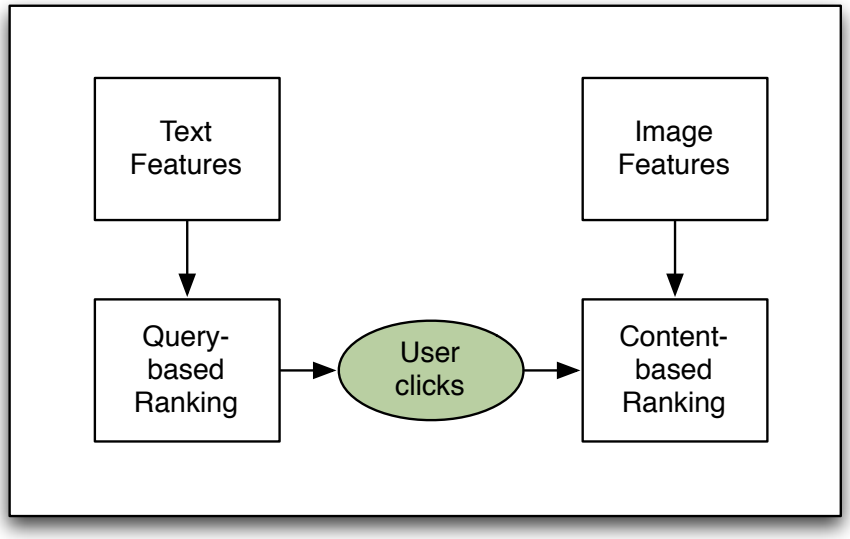


Figure 22: A key distinction of our proposed approach is the use of the query logs of one type of image retrieval system (text-query based) as training data for another type of image retrieval system (content-based).

feasibility of using click-data in an open retrieval system to measure image similarity. In particular, we propose to derive relative comparisons from the aggregated co-click statistics and the average position of the images in the search results.

This chapter is divided into three parts. Section 3.2 introduces related works in learning distance functions for Web image retrieval. Section 3.3 proposes methods to derive measurement of image similarity from text-based image search query logs, and use such information as training data. Section 3.4 introduces methods to learn query-specific distance functions.

3.2 Related works

Our work is mainly related to two areas of research. The first area is related to distance learning research in machine learning, and the second is related to exploring log data as relevance feedback in Web search. This work briefly reviews some representative works in both areas.

Learning distances

In spite of the observation [83] that human perception sometimes can not satisfy triangular equality, distance metrics, such as Euclidean distance, have been used extensively in

large-scale Web image retrieval systems for its simplicity and efficiency [96, 23]. Methods to improve the accuracy of Euclidean distance has been proposed previously, including unsupervised learning techniques such as Metric Multidimensional Scaling [19], Locally-Linear Embedding [84], Isomap [102], Pyramid Match Kernel [34, 59], and supervised distance learning techniques [113, 91, 108, 33, 115] that learn a *weighted* Euclidean distance function.

Previous works to learn weighted Euclidean distance functions differ in how training samples are collected and how they formulate the optimization. For example, Xing et al. [113] learns a weight vector that minimizes the number of violated constraints in the training data, structured in the form of pairwise comparisons ("A and B are similar"). Schultz [91] adopted an optimizing approach that is analogous to a soft-margin SVM in that the relative comparison ("A is more similar to B than A to C"). In these approaches, a single distance function is learned and used to compare all images in the database. This work studies metric learning in the context of hybrid image retrieval system, and proposes to adopt the learning approach in [91] to learn query-specific distance functions.

Our work is mostly related to exemplar-specific distance learning approach proposed by Frome et al. [33]. Learning distance per image, however, does not scale for Web image retrieval due to the high cost of storing the weights for each image. From the observation that search queries term frequency usually follows power law distribution [89], our proposed approach can be applied to images retrieved by the most popular queries with modest additional cost to store the feature weights.

Use of search engine logs

The use of logging data as a form of relevance feedback [119, 96, 88, 86] has been explored previously by Web information retrieval and content-based image retrieval communities [51, 52, 91, 104, 45, 78, 39, 10, 29]. Joachims [51] conjectured that click-through statistics often convey judgment of document relevance with respect to the query, and confirmed this hypothesis with an eye-tracking study [52]. Uchihashi et al. [104] proposed a *content-free* image retrieval system entirely based on modeling the click statistics of the image retrieval

systems through collaborative filtering [39] techniques. Radlinski et al. [78] demonstrated that click-through data is not reliable for deriving absolute relevance judgment as it is affected by the retrieval quality of the underlying system, but relative comparisons ("A is more relevant to the query than B") are reasonably accurate. Schultz et al. [91] proposed to learn a ranking function from co-click statistics for Web document retrieval, and Jain et al. [45] applied similar techniques to the retrieval of Web images.

Our work is also related to the work by Hoi et al. [41], who demonstrated that the click patterns made by users of content-based image retrieval system can be used as relevance feedback signals to refine image distance function. This work has several key differences comparing with [41]. First, training data used in this work is in the form of relative comparison, as opposed to "relevant" or "irrelevant" labels. Second, this work proposes to learn a unique distance function for each query, as opposed to a global distance for all images. Third, this work derives training data from *text-based* image retrieval system. This approach allows one to leverage large quantities of feedback data from popular incumbent Web retrieval system, and use it to improve the performance of new system that may not have sufficient relevance feedback data of its own.

3.3 Measure Image Similarity with Co-click Statistics

A search session [64] starts when the user initiates an image search task (perhaps by typing the URL of a commercial search engine), and ends when the user leaves the search engine, or no longer actively searches on the site. During this time, users usually have viewed a large set of images, and may have *clicked* on one or more images that satisfy his or her search criteria. Such browsing behaviors are recorded as a part of the image search engine query logs. In a single image search session, if image x_i and image x_j are both clicked by the user, we say they are co-clicked.

This work studies whether two images that are co-clicked more often are similar to each other than to a third image co-clicked less often. The intuition is that when conducting an image retrieval task, many users have a pre-determined mental image of what they are looking for. Therefore, during the process of browsing through the search results, users may

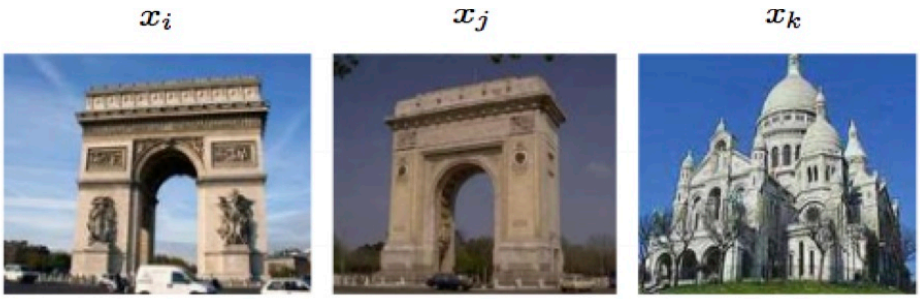


Figure 23: Image x_j is more similar to query image x_i than image x_k is to x_i.

conduct an implicit comparison between the images retrieved with the target image. Only images similar to the target image are selected while others are seen but ignored. Therefore if we aggregate the co-click statistics over all search sessions conducted within a sufficient period of time, then images that are clicked more often are more similar to each other. Our goal is to derive reliable measurement of image similarity from such aggregated co-click statistics, and use it to train query-specific distances.

One can imagine several situations when such hypothesis is not true. For example, a person may not have concrete search criteria (e.g. casual browsing) or the search criteria may change over time. In this case, the images clicked may not exhibit any semantic or visual relationships at all. The hope is that by aggregating the query sessions made by billions of Web users, the distinctive click patterns may emerge to capture how majority of the people perceive image similarities.

3.3.1 Image comparison with co-click statistics

In this work, we propose to derive relative comparisons ("image A is more similar to image B than A is to C") from co-click statistics, and use such information to learn query-specific distance functions. Comparing with pairwise comparisons (e.g. "image A is similar/dissimilar to B"), relative comparisons is context-dependent, and contains richer set of information that can be used to derive the relative ordering of the images. For example, given the three images shown Figure 23, although both image x_j and x_k are related to the image x_i (e.g. all Paris landmarks), most would agree that x_j is more similar to x_i than x_k is to x_i.

One can use co-click statistics as absolute and quantitative measurements of pairwise

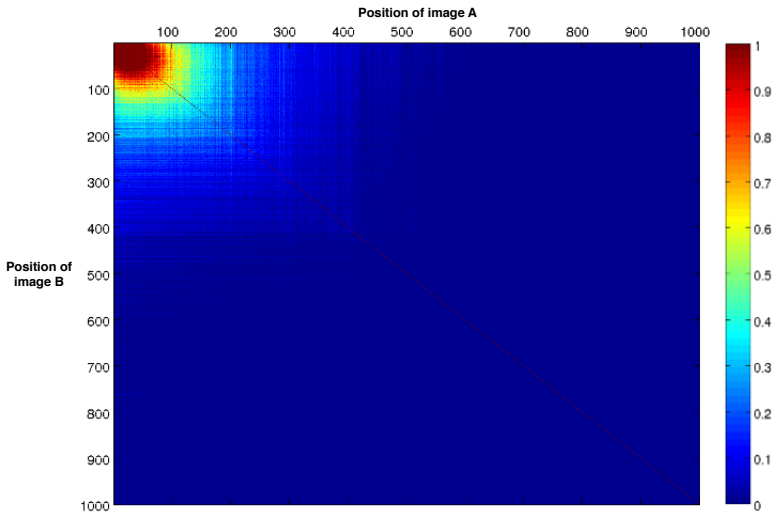


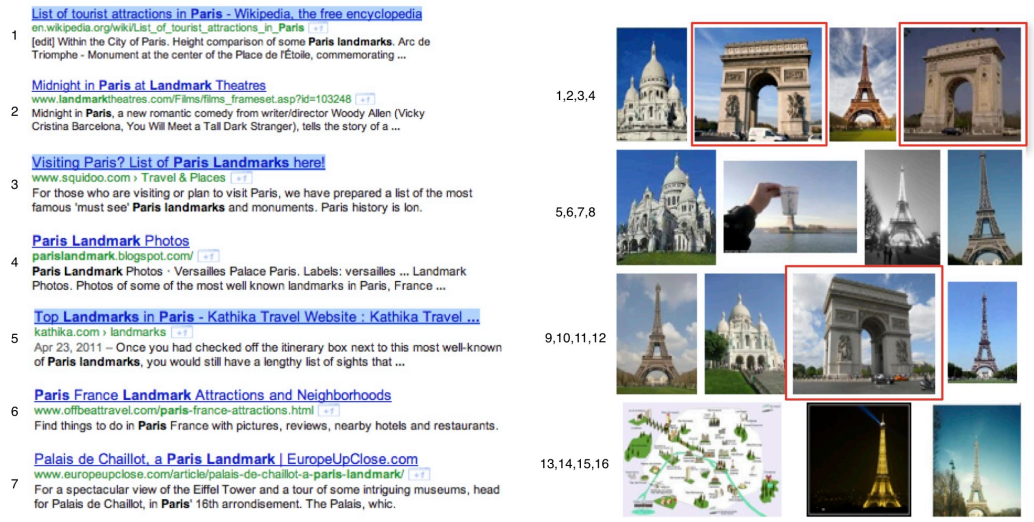
Figure 24: The correlation between co-clicks between two images and their respective position in the search results. The point (x, y) on the two dimensional plot (x, y) represents the average amount of co-clicks received by images with x and y as their respective position in the search results. On average, the likelihood of a user click on an image tends to decreases as the rank increases.

similarity, and use it to compare or rank images. For example, given a query image x_i and two candidate images x_j and x_k, one can determine which candidate image is more similar to the query image with the following equation:

$$\delta(x_i, x_j, x_k) = \left\{ \begin{array}{ll} x_j, & \text{if } C(x_i, x_j) > C(x_i, x_k) \\ x_k, & otherwise \end{array} \right\} \tag{5}$$

where $C(x_i, x_j)$ is the number of search sessions where image x_i and x_j are co-clicked. Adopting Equation 5 for image comparison assumes that we have accurate measurement of pairwise distances (or similarity). However, as this work proposes to derive image similarity from user click-patterns, the order in which images are presented to the user can significantly affect whether the likelihood of images being clicked by the users.

Figure 24 illustrates the position bias by showing the correlation between co-clicks between two images and their respective position in the search results. The point (x, y) on the two dimensional plot (x, y) represents the average amount of co-clicks received by images with x and y as their respective position in the search results given a set of popular queries. On average, the likelihood of a user click on an image tends to decreases as the rank increases. Such position bias is due to the well document tendency [51, 20] for search engine



(a) Web search (b) Image search

Figure 25: An example of Web/Image search results. Web documents or images that are clicked on by the user during a search session are highlighted. We can reasonably expect that those images (or documents) ranked ahead of the clicked images (or documents) are observed but not clicked.

users to exist search when the first relevant image is found, regardless of whether there is a more or equally relevant images positioned further down in the list of search results.

To address this problem, we proposes to incorporate the *average position* of the images into the comparison function. This is based on the observation that, when a ranked list of Web documents are presented to the Web search engine users, documents that are clicked on are more semantically relevant to the query than those that are observed but not clicked on [51]. In the absence of information on what documents users have observed, a commonly used assumption is that user examine search results sequentially and therefore all the documents ranked ahead of the last clicked image is considered observed. For example, in Figure 25(a), the documents that are clicked are highlighted. One can reasonably expect that document 2 and 4 are observed but not clicked.

We extend this intuition to the domain of image search: images that are clicked more frequently are more similar to each other than those ranked higher but clicked less frequently. The process of labeling image triplets contains the following two steps: first, we count the number of search sessions where a pair of images is co-clicked, denoted as $C(x_i, x_j)$ for image x_i and x_j. Note that $C(x_i, x_j)$ is aggregated over all possible queries these images can be

retrieved with to generate sufficient sampling of co-click statistics. Next, we computed the average position of each image relative to other images in the same query during the time the data is collected. In this work, we refer the average position for the image x_i given query q as $P_q(x_i)$, where $P_q(x_i) < P_q(x_j)$ when x_i is ranked ahead of x_j. Note that the position of the images are query-dependent.

The resulting relative comparison function is shown below,

$$\delta(x_i, x_j, x_k) = \left\{ \begin{array}{ll} x_j, & \text{if} \quad C(x_i, x_j) > C(x_i, x_k), \quad P_q(x_k) < P_q(x_j) \\ x_k, & \text{if} \quad C(x_i, x_j) < C(x_i, x_k), \quad P_q(x_j) < P_q(x_k) \\ \emptyset, \text{otherwise} & \end{array} \right\} \tag{6}$$

If the position-constraints is not satisfied, then the function will output $\emptyset$, indicating that we do not have sufficient information to determine which of the candidate images is similar to the query image. We apply Equation 6 to all permutations of image triplets x_i, x_j, x_k sampled from the images produced by query q. We remove the triplets when image comparison cannot be reliably estimated from the co-click statistics (labeled with $\emptyset$).

One can also combine co-click statistics with other types of distances, such as Euclidean distance (L2) derived from image features, using the following equation:

$$\delta(x_i, x_j, x_k) = \left\{ \begin{array}{ll} x_j, & \text{if} \quad C(x_i, x_j) + d_2(x_i, x_k) > C(x_i, x_k) + d_2(x_i, x_j), \quad P_q(x_k) < P_q(x_j) \\ & \text{or} \quad d_2(x_i, x_k) > d_2(x_i, x_j), P_q(x_k) > P_q(x_j) \\ x_k, & \text{otherwise} \end{array} \right\} \tag{7}$$

where $d_2(x_i, x_k)$ is L2 distance computed over image features. Comparing with Equation 6, Equation 7 combines co-click statistics with the distances produced using L2 distance over image features when the rank constraint is satisfied, otherwise only L2 distance is used. In this case, proper distance scaling between d_2 and C (such as those used in [33]) is needed.

As Web search engines typically do not have control over a user may interpret the search results and interact with the retrieval system, it is possible that the images clicked may not exhibit any semantic or visual relationship at all in some search sessions. Our hope is that despite the subjectivity in human perception of image similarity, one can still find distinctive click patterns for subsets of queries and images that capture how majority of

the people perceive image similarity. For this reason, instead of considering image triplets generated from each search session as a separate measurement of image similarity, we use the aggregated statistics over 1-year worth of image search query logs.

3.4 Learning query-specific distance for Web image search

Our goal is to learn a weighted Euclidean distance d_{w_q} for each query q. Each query is associated with a set of images X_q and $x_i \in X_q$ is represented by a M dimensional feature vector $\vec{x_i} = \{\vec{x_i}^1, ..., \vec{x_i}^M\}$. We define the query-specific Euclidean distance between image x_i and x_j

$$d_{W_q}(\vec{x_i}, \vec{x_j}) = \sqrt{\sum_{m=1}^{M} \vec{W_q}^m (\vec{x_i}^m - \vec{x_j}^m)^2}. \quad (8)$$

where W_q is a M dimensional weight vector over the features.

Given a training set T_{train} of n relative comparisons, our goal is to learn the weight vector W_q over the features such that the training error (i.e. the number of violated constraints) is minimized. Using the training image triplet shown in Figure 23 as an example: as it is clear that image x_j should be considered more similar to image x_i than x_k is to x_i, the learning goal is to find distances between images such that relationships of this type holds, for example, that the distance $d_{W_q}(\vec{x_i}, \vec{x_j}) < d_{W_q}(\vec{x_i}, \vec{x_k})$. Of course, if all our images are from the training set, then we don't need the distance functions at all; we can simply rank images based on the comparison based on co-click statistics. However, such supervised information is typically may not be available for all image triplets in the database (due to the position bias), and certainly not available for new images.

Following [91], finding a solution of minimal training error is equivalent to finding a W_q that fulfills the following constraint.

$$\forall (i, j, k) \in T_{train} : d_{W_q}(\vec{x_i}, \vec{x_k}) - d_{W_q}(\vec{x_i}, \vec{x_j}) > 0. \quad (9)$$

As the solutions is typically not unique, learning methods have been proposed to select W such that the learned distance remains as close to an un-weighted Euclidean distance as possible. Following [91] we adopt the max margin framework that minimizes the norm of

weight vector $\vec{w_q}$, this leads to the following optimization problem:

$$min\frac{1}{2}||\vec{w_q}||^2 \tag{10}$$
$$s.t. \quad \vec{w_q} \cdot (\vec{\Delta}^{x_i,x_k} - \vec{\Delta}^{x_i,x_j}) > 0$$
$$\forall (i,j,k) \in P_{train} \tag{11}$$
$$\vec{w_q}^m \geq 0 \quad \forall m \in \{1,...,M\}$$

where $\vec{\Delta}^{x_i,x_k} = (\vec{x_i} - \vec{x_k})^T(\vec{x_i} - \vec{x_k})$. Compared with standard quadratic programming such as SVM, this optimization has an additional constrain on $\vec{w_q}$, which needs to be positive such that it meets triangle inequality of distance. We add slack variables ξ_{ijk} to each triplet to account for constrains that cannot be satisfied, we then get the following optimization problem:

$$min\frac{1}{2}||\vec{w_q}||^2 + C\sum_{i,j,k}\xi_{ijk} \tag{12}$$
$$s.t. \quad \vec{w_q} \cdot (\vec{\Delta}^{x_i,x_k} - \vec{\Delta}^{x_i,x_j}) > 1 - \xi_{ijk}$$
$$\forall (i,j,k) \in P_{train}$$
$$\xi_{ijk} \geq 0, \quad \vec{w_q}^m \geq 0 \quad \forall m \in \{1,...,M\}$$

where the scalar C is the trade-off parameter between the empirical loss term and the regularization term. The form in Equation 12 is similar to the soft-margin SVM [77]. We solve this optimization problem using sub-gradient method based on [99]. This method does not directly depend on the number of training samples and is very fast in practice.

CHAPTER IV

EVALUATION WITH HUMAN LABELED DATA-SETS

Chapter 3 introduced learning query-specific distance functions from click-patterns derived form image search query logs. This chapter evaluates the accuracy of learned distance functions with manually labeled data-sets. Specifically, we collected a large set of ground-truth data by having human raters to judge the relative similarities among images. Our results demonstrate that query-specific distance functions learned with co-click statistics outperform Euclidean distance and a query-independent distance function learned from the same training data.

Note that due to the added degrees of freedom, using query-specific distance functions will, in many situations (especially when training and testing data are drawn from the same underlying distribution), always produce more accurate image comparisons than query-independent distances. Therefore, this experiment is conducted in a way such that evaluation method (e.g. perceptual comparison test) are not directly tied to how training data is produced (e.g. co-click statistics in Web image search).

4.1 Introduction

Various approaches to evaluate distance functions have been used in the past. One such approach is to manually assign test images with class labels, and then apply distance functions (as a part of a K-Nearest Neighbor classifier) to classification tasks [33, 117, 59]. For example, Frome et al. [33] applied the learned distance function to the task of object recognition, with experiments conducted using images and associated class labels from the Caltech-256 dataset [36]. An alternative evaluation approach commonly used by information retrieval community is to label each test image as either "relevant" or "irrelevant" to the query image based on their shared text annotation [98, 42], and compute precision and recall scores from the search results produced by the distance function in question.

The drawback of using class labels or image annotations is that text information is

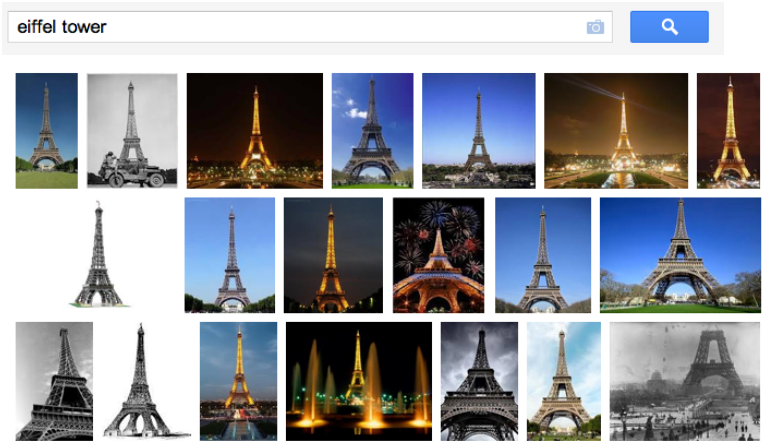


Figure 26: Images retrieved from Google images with query "Eiffel Tower." Although the image in the search results all share the same query, some images are more similar to each other than others.

often not sufficiently descriptive of the images. For example, Figure 26 shows the search results produced with the query "Eiffel Tower:" although these images are annotated with the same query, it is evident that some images are more similar to each other than others. Therefore, evaluating distance functions for the task of re-ranking search results requires more descriptive and precisely labels.

This work adopts an evaluation approach that first asks human raters to compare sets of images, and apply the resulting human ratings as labels to evaluate a particular distance function. Such perceptual similarity experiments have been proposed previously [17, 92, 82, 69] to evaluate content-based image retrieval systems. Typically human raters are asked to compare sets of images and assign either an quantitative similarity score to a pair of images (e.g. image A and B are very similar) or qualitative and relative comparisons (e.g. image A is more similar to image B than image A to C).

This work uses a variation of relative comparison test as shown in Figure 27. The query image is displayed at the top and two candidate images are displayed at the bottom of the screen. Human raters are instructed to indicate which of the two candidate images is more similar to the query image. As candidate images can all be similar or dissimilar to the query image, the experiment also allows users to select *cannot decide*, Also, as the perception of image similarity can be subjective with respect to the experiences of the raters, we assign

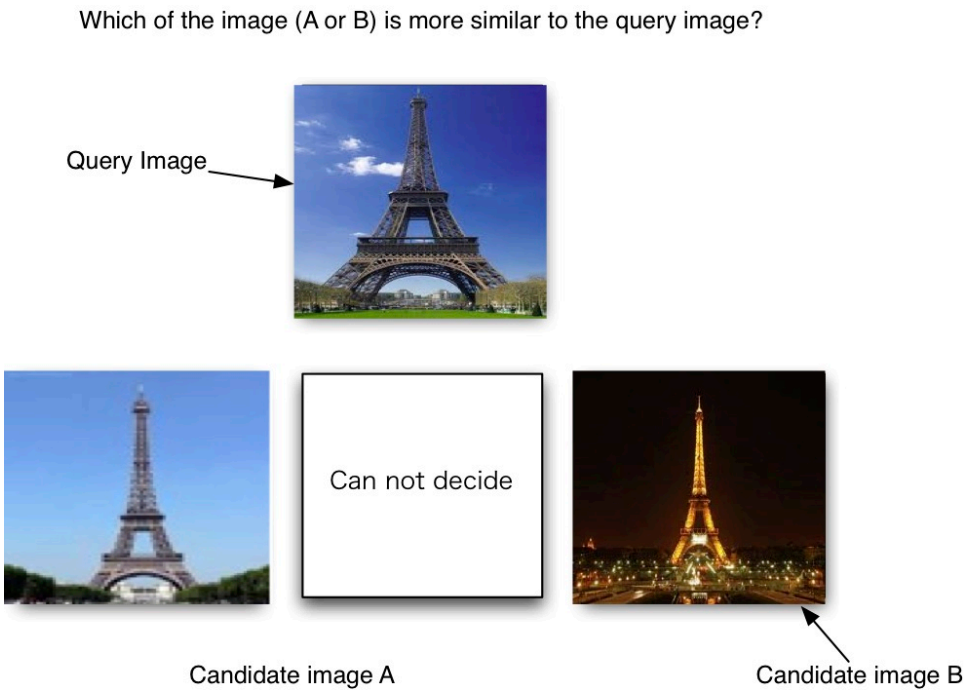


Figure 27: The triplet rating interface with example images. Three images are displayed to the user. The query image is displayed at the top, while the candidate images are displayed below the query image. If the user consider candidate image A to be more similar to the query image than candidate image B, then the user is instructed to click on image A and the response is recorded, and vice versa. If the decision is difficult to make (i.e. both candidate images are similar or dissimilar to the query image), then the user can click on the center button to indicate the lack of any difference.

each sets of images to multiple raters, and only consider images with consistent label from all raters. Therefore, we propose to measure the accuracy of a distance function by comparing the its output with rater selection on the testing images.

The rest of the chapter is divided into three parts: Section 4.2 presents previous works related to measuring human perceptual similarities, section 4.3 presents the detailed experiment methodology and Section 4.4 presents the experiment results.

4.2 Related works

The use of human judgment to measure image similarity, including both the absolute and relative rating scales, has been proposed in the evaluation of PicHunter system [17]. Three types of user experiment were proposed (with the interface showing in Figure 28), including a) absolute similarity test, where users were asked to indicate the similarity between two

images on a 5 point scale; b) relative comparison test, where users were asked to judge the degree of similarity between the query image and two test images on a rating scale, and c) Two-alternative forced-choice (2AFC), a variation of relative comparison that forces users to choose between two ratings (left, right). Studies have found that absolute and relative comparison tests are highly correlated with each other [101]. An alternative method to measure perceptual similarity is the *table scaling* [82] method, where users were asked to arrange a set of images on a table (or computer screen) so that distances between them were inversely proportional to their perceived similarity.

The experiment methodology shown in Figure 27 is a variation of relative comparison test. Relative comparison test has the advantage that it is more objective and independent of any criteria that the rater has to apply in rating tasks, and with 2AFC test, choosing one of the two images was generally considered to be an easy task [69] by the raters and more tasks can be completed given a fixed amount of time. Our methodology is analogous to relative comparison test with 3 point scale (-1, 0, 1) [1]. Such configuration makes the rating task easier by reducing the range of decisions, but still maintain the "can not decide" option to avoid rating noise as a result of forced choice.

4.3 Experiment methodology

4.3.1 Sampling queries from image search logs

We selected queries belonging to four categories of visual concepts, *person*, *product*, *animals* and *places*, as these categories contain many of the most frequent terms people use to query commercial image search engines. Also, as such categories usually have distinctive visual appearances, they are commonly used as a part of benchmark database for evaluation of recognition systems [35, 25]. We also included a fifth category referred as *polysemy*, which are queries with multiple semantic and visual concepts. For example, the query *apple* produces images related to both the company and the fruit. We conjecture that ranking based on image similarity is mostly helpful on the search results retrieved with such ambiguous queries.

[1] It is also analogous to 2AFC test with a choice to "skip" a particular test case.

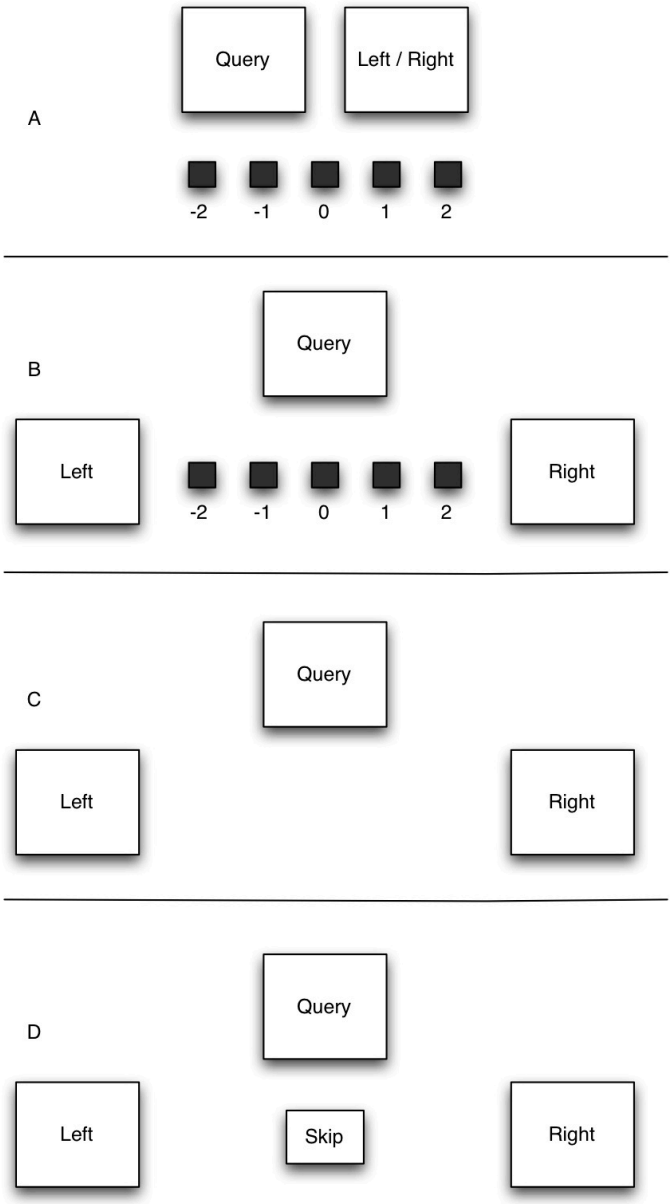


Figure 28: Types of experiments to measure image similarity: a) the absolute-similarity configuration, b) the relative-similarity configuration with 5 point scale, c) 2AFC configuration, d) the two-choice relative comparison test we use

Table 6: A list of 50 (44 unique) queries were sampled from a set of 10,000 most popular queries on Google image search.

People	lady gaga, steve jobs, bill gates, barrack obama, brad pitt, taylor swift, kobe bryant, david beckham, paris hilton, allen iverson
Product	iphone, bmw z8, ipod, coca cola, nokia phone, electric guitar, dell computer, zune, paper clips, alarm clock
Animal	tiger, fish, cat, dog, pig, beetle, zebra, chicken, bird, jaguar
Places	golden gate, eiffel tower, stanford university, beach, fuji, great wall, washington, store, lincoln memorial, notre dame
Polysemy	apple, tiger, jaguar, washington, notre dame, cup, fuji, beetle, darwin, crane

We selected 10 queries for each category with the following selection methodology: first, we collected 10,000 of the most frequently searched for queries on Google images during the month of July 2010; second, we uniformly sample queries from this list, and manually assign each query to one of the five categories illustrated above. A query is removed from consideration if it does not fall into any of the five category, or if the retrieved images contain pornography or other inappropriate content. This process is repeated until each category contains 10 queries. The complete list of queries are shown in Table 6.

4.3.2 Sampling image triplets from search results

To evaluate query-specific distance functions, we propose to sample image triplets from the retrieved images and have them labeled by human raters. This section describes the methodology used to sample image triplets from the top image produced by a query.

Given each query, we extracted the top 100 search results from Google image search, with the strict safe search filter. The top 100 images are used instead of the 1000 images available to us due the following two considerations: first, we observed that the relevance between the retrieved images and text query degrades significantly beyond the top 100 results retrieved from Google images; second, as users usually follow the order in which search results are presented to them, they are more likely to select a query image from the top search results to conduct hybrid image retrieval.

We randomly sampled 25 testing images from the top 100 search results. Since 2300

unique combinations image triplets (25-choose-3) can be sampled from the testing images, and each of image in a triplet can be the query image, there are 6900 possible testing image triplets. We randomly sampled 1000 (14.5% of 6900) image triplets and have them labeled by the human raters.

4.3.3 Experiment User Interface and Procedure

The interface is shown in Figure 27. The resolution of 1920x1080 is used. Three images are displayed to the user. The query image is displayed at the top, while the candidate images are displayed below the query image. User has the option to select either left or right candidate image, or select "can not decide" to skip this task. After a selection is made, a new task is displayed to the user. Standard 27 inch monitor will be used, where the browser (chrome) is maximized to occupy the entire screen. When multiple systems are used to conduct studies, the type of monitor, browser, mouse and other interfaces are identical to each other.

For each query, we partition the 1000 triplets into 20 triplet groups, and each group is presented to three different human raters for labeling. 7 raters participated in this experiment, each rater spent 4 hours a day (50 minutes rating-time with 10 minutes rest-time) for a total of 10 business days. We only consider testing images that received consistent labels from the three raters.

4.4 Experiment Results

This section presents a set of experiments designed to evaluate the quality of the relative comparisons generated from the query logs and the accuracies of distance learned from such information. The testing images are collected with the procedure listed in Section 4.3.2. We used **Google-L2** distance function as benchmark for comparison. Google-L2 distance is a highly optimized distance function over the image features used by Google Similar Images [1].

Figure 29: A sample of testing triplets where comparison results derived from co-click statistics disagree with those based on applying Euclidean distance over the image features. Each numbered row represents a testing triplet. For each triplet, the 1st image is the query image and the 2nd and 3rd are candidate images. The candidate images are arranged such that the 2nd image is more similar to the query image based Google-L2 distance over image features, and the 3rd image is more similar based on co-click statistics.

4.4.1 Analysis 1: Examples of Results

Figure 29 contains a sample of testing triplets. In particular, it contains triplets such that the comparison decision derived from co-click statistics disagrees with those derived by applying Euclidean distance over the image features. Each numbered row represents a testing triplet. For each triplet, the first image is the query image and the second and third image are candidate images. The candidate images are arranged such that the second image is more similar to the query image based on the **Google-L2**, and the third image is more similar based on **co-click** statistics using Equation 6 in Chapter 3.

We observe that **Google-L2** distances are sufficiently accurate when two images contain the same objects (row 1, 2, 3, 5, 20) or share dominant visual cues (4, 6, 7). Co-click

similarities are less accurate in such cases – indicating that images clicked during a search session are likely to be semantically and visually similar only up to a point. Images that are duplicate or near-duplicate of each other are not necessarily the most frequently clicked pair during a search session. On the other hand, when a particular visual concept (such as apple logo) has high intra-class variance with respect to the image features, co-click statistics tends to be more accurately capture the semantic similarity among the images (8 - 15, 16, 17, 18, 19, 21, 22, 23). It is our hope that by learning feature weights from the co-click statistics, those most discriminative features in this query, such as shape of the logo, are likely to have more weights over other features (color, etc).

4.4.2 Analysis 2: Accuracy of Co-click statistics

Figure 30 compares the accuracy of co-click statistics with other measurement of image similarity such as Euclidean distance over image features. The ROC curve is computed by adjusting the threshold of the distance comparison. **Google-L2** represents the highly optimized distance function used by Google Similar Images. **Coclick** represents comparison with Equation 6 in Chapter 3; and **Coclick+L2** compares images using a combination of co-click statistics and *Google-L2* distance derived from image features using Equation 7 in Chapter 3. We scaled scale the two distances so that they have the same variance with approach similar to [33]. To ensure that one can fairly compare method with three outcomes (e.g. $x_j, x_k, \emptyset$) with those with two outcomes (e.g. x_j, x_k), we removed all the testing data where the output of co-clicks resulted in $\emptyset$ from evaluation.

Figure 30 shows that when used separately, for all categories of queries other than polysemy, **Google-L2** distance is more accurate than co-click statistics. There are three likely reasons for such results: first, **Google-L2** distance, used by Google Similar images, is highly optimized over the image features; Second, the perceptual similarity test we used to obtain the labels ("which of the image is more similar to the query image") is conducted without giving raters a specific search task in mind. For this reason, raters are more likely to base their judgement on what they perceive as the most dominant visual properties of the images (e.g. dominant color of the background, etc).

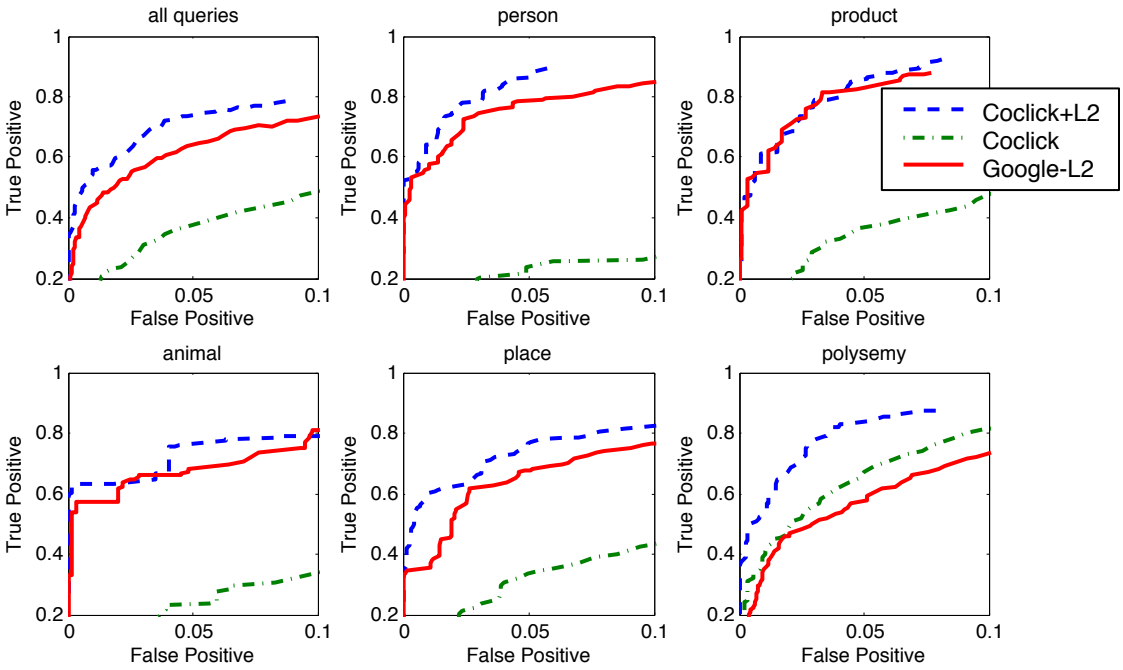


Figure 30: The accuracy of co-click statistics in predicting user comparison ratings. We compute the average true positive and false positive rate under various distances for each categories of queries. **Google-L2** represents the distance function used by Google Similar images; **Coclick** represents co-click statistics (equation 6 in Chapter 3); and **Coclick+L2** combines co-click statistics with distances over image features with Equation 7 in Chapter 3.

Third, as shown in column 1 of Figure 29, images that are near-duplicate of each other are frequently shown in the top search results (especially in queries associated with *product* categories) and therefore in the testing image triplets. In such cases, when one of the near-duplicate image is used as the query image, raters often choose the other one as the more-similar image. Such near-duplicate images can be easily identified by using global features. However, we observed that near-duplicate images do not usually receive more clicks than images that are visually less similar but semantically related. We conjecture that this is due to the fact that when searching for photo from the image search results, search engine users are unlikely to exam near-duplicate images during the same search sessions. For this reason, such task-dependent click patterns has different properties as those derived from perceptual similarity test.

Figure 30 also shows that for queries in *polysemy* categories. **Co-click** statistics outperforms **Google-L2** distances. This is due to the observation that, as shown in column 2 and 3 of Figure 29, a particular visual concept (such as apple logo) has high intra-class variance with respect to the image features. Images can be similar in multiple feature dimensions such as color (green apple, green logo) or shape. Query-independent distance functions, such as **Google-L2**, has limited capacity to select features that are important to disambiguate images produced by the the text-query. For this reason, **Google-L2** is less accurate than when co-click statistics is used. By learning feature weights from the co-click statistics, those most discriminative features in this query, such as shape of the logo, are likely to have more weights over other features (color, etc).

Figure 30 also shows that combining both co-click statistics and the Euclidean distance over the image features (**Coclick+L2**) produces more accurate estimation of image distances than when each is used separately. This result is not surprising as combining two largely independent sources of information usually produce more accurate results than when either one is used separately.

Figure 31 presents more detailed results of image comparison using Equation 6, itemized for each query. **TP/FP** represents true positive/false positive rates – the percentage of testing triplets where the label agrees/disagrees with the output of Equation 6. $\emptyset$ represents

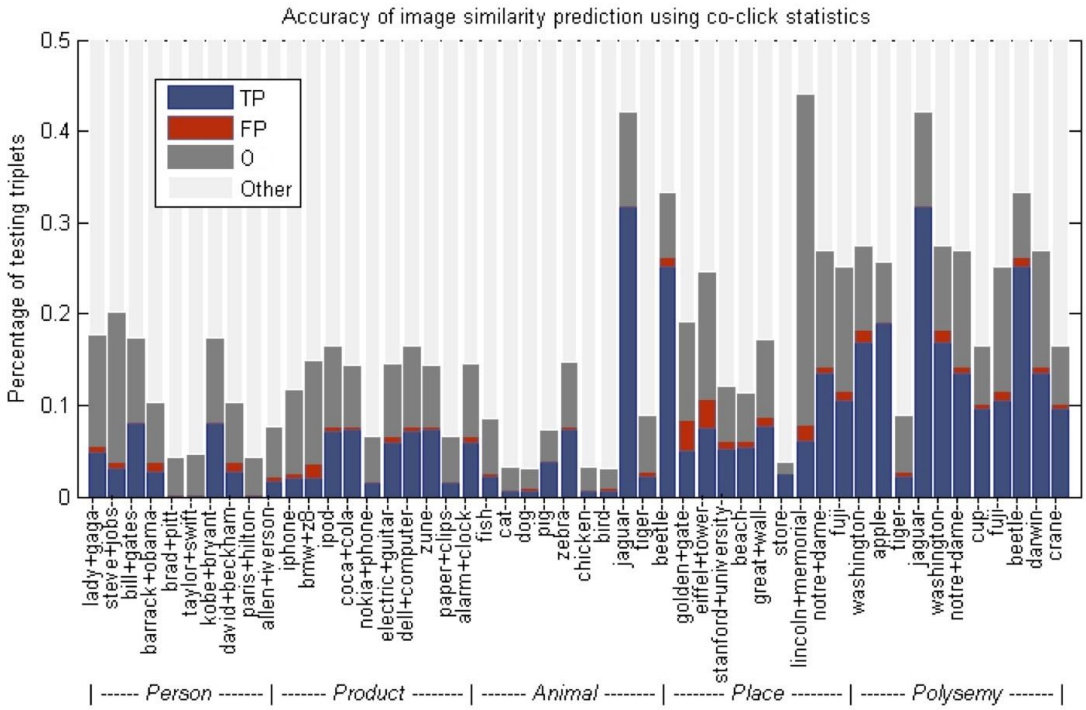


Figure 31: The accuracy of co-click statistics. **TP/FP** represents true positive/false positive rates – the percentage of testing triplets where the label agrees/disagrees with the output of Equation 6. ∅ represents testing triplets where rank constraints are not satisfied. **Other** represents testing triplets with missing or inconsistent labels from different raters.

testing triplets where rank constraints are not satisfied. **Other** represents testing triplets with missing or inconsistent labels from different raters. Figure 31 shows that for all queries, majority (more than 50%) of the image triplets received the rating of **Other**, more so in categories such as person, product and animal than polysemy. It also shows that co-click statistics along can not accurately compare images.

4.4.3 Analysis 3: Accuracy of query-specific distance

This section describes a set of experiments designed to evaluate the accuracy of learned query-specific distance functions, and compare it against the global (query-independent) [91] distance functions. We sample training triplets from the combined distance using Equation 7, and use it to train query-specific distance function. We represent image features as a fixed dimension feature vector derived from first concatenating and quantizing various types of image features such as color, texton, and wavelets, and use kPCA with Histogram intersection [100] kernel is used to reduce the dimensionality of the feature space. The top

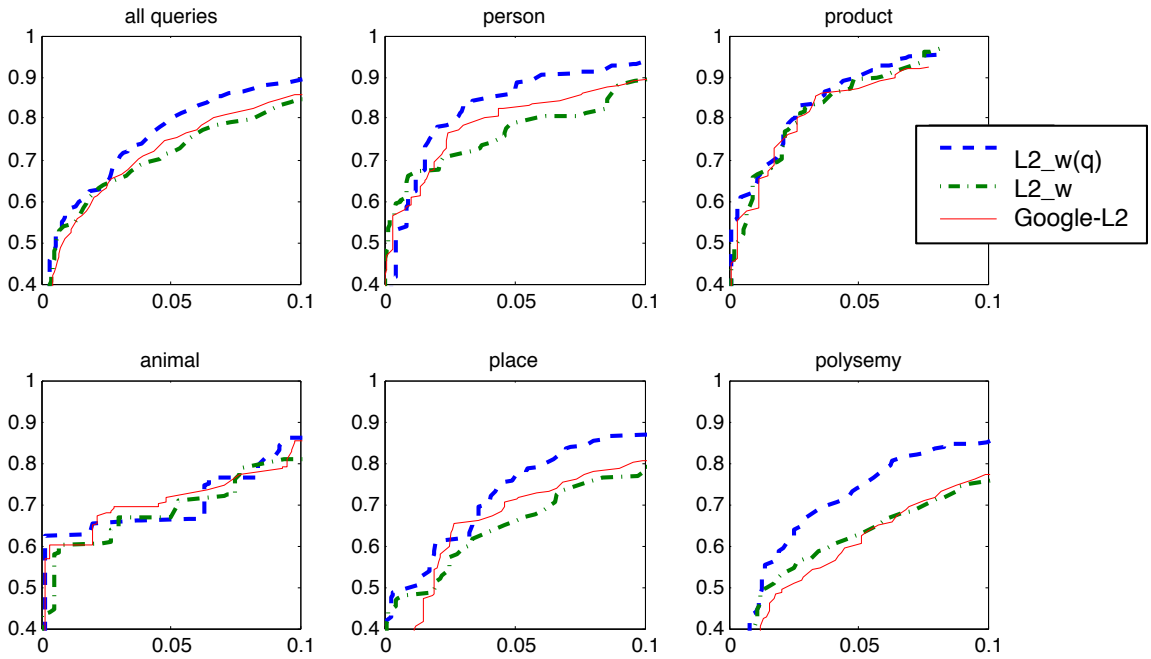
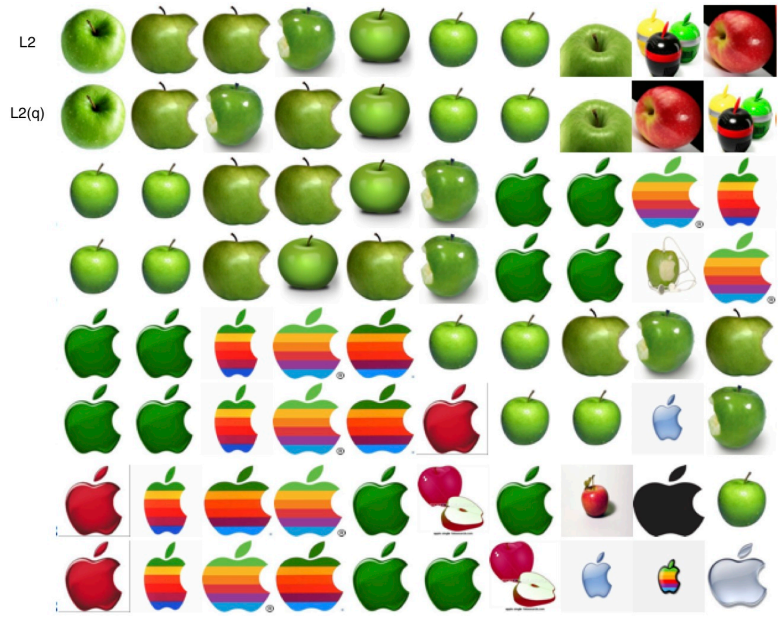


Figure 32: The accuracy of query-dependent distances. We compute the average true positive and false positive rate under various distances for each categories of queries. **Google-L2** represents the highly optimized distance function Google Similar images currently uses; $\mathbf{L2}_w$ represents query-independent distance learned from co-click statistics. $\mathbf{L2}_{w_q}$ represents query-dependent distance. We observe that query-dependent distance is more accurate than the two competing methods, especially for images related to polysemous queries.

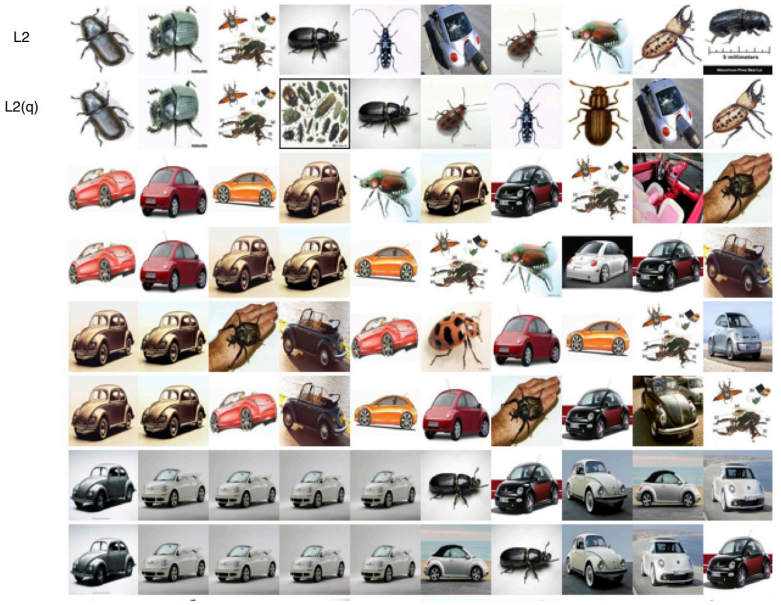
59 dimension is used in this work.

Figure 32 shows the accuracy of query-specific distances. We compute the average true positive and false positive rate under various distances for each categories of queries. $\mathbf{L2}_{w_q}$ represents query-dependent distance; $\mathbf{L2}_w$ represents query-independent distance learned from the same training data. Figure 32 shows that $\mathbf{L2}_{w_q}$ outperforms both $\mathbf{L2}_w$ and **Google-L2**. The improvement is more significant in category *person, place* and *polysemy*. Note that learning a single query-independent distance function resulted in worse performance than query-specific distance functions.

Figure 33 presents a set of ranking results where the learning query-specific distance is particularly beneficial. Each row presents the top 10 nearest neighbor images retrieved with the first image as the query image. The odd number of rows (1, 3, 5, ...) are ranking



(a) apple



(b) beetle

Figure 33: Examples of image ranking results. Each row presents the top 10 nearest neighbor images retrieved given the first image as the query image. The odd number of rows (1, 3, 5, ...) are ranking based on **Google-L2**, while the even number of rows (2, 4, 6, ...) are those base on query-dependent distance $\mathbf{L2}_{w_q}$.

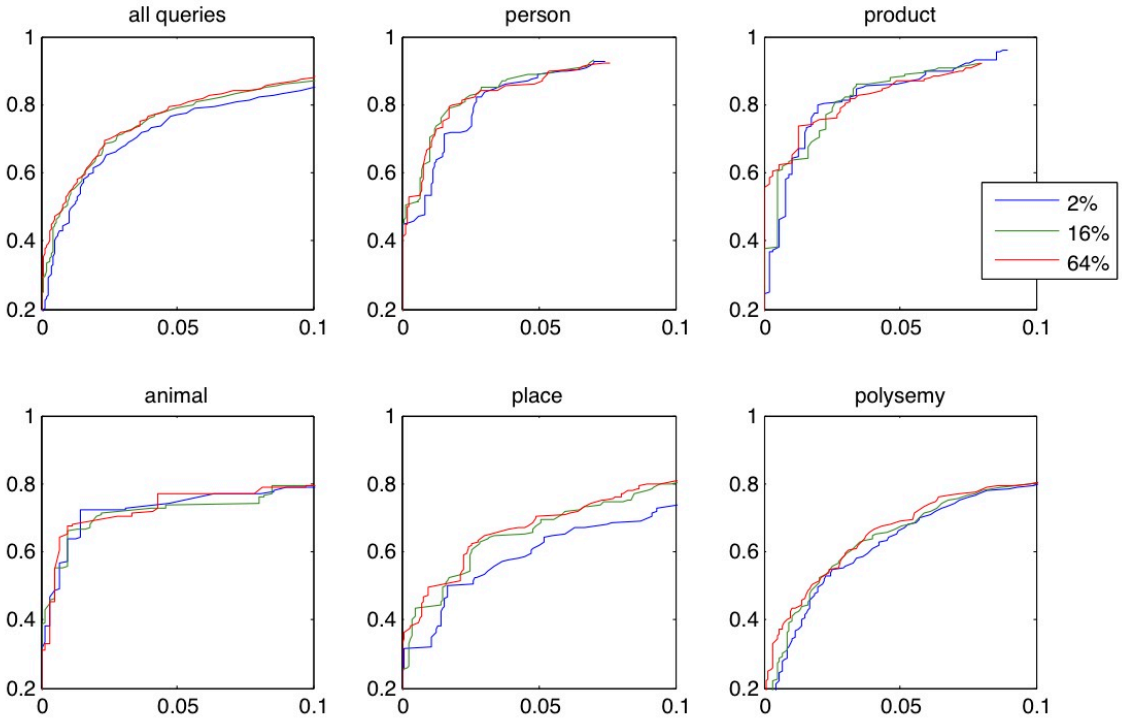


Figure 34: The accuracy of query-specific distance given the number of available training triplets.

based on **Google-L2**, while the even number of rows (2, 4, 6, ...) are those base on query-dependent distance ($\mathbf{L2}_{w_q}$).

4.4.4 Analysis 4: Size of training data on accuracy

As it is computationally expensive to train with all available training triplets, this section studies the effect of number of training data on the accuracy of the learned distance. In our experiment, we obtained an average of 2 million training triplets for each query. Figure 34 presents the accuracy of query-specific distance given the number of available training triplets. We randomly sampled **2%**, **16%**, and **64%**, of the triplets from the all available training data. The result shows that the testing accuracy improves quickly as we increase the number of training data from 2% to 16%. The accuracy distance functions trained from 15% of the available data is comparable with those trained from all the data.

4.5 Conclusion

We demonstrate that co-click statistics derived from text-based search engine query logs can be used to predict how human will compare images based on perceptual similarity. We demonstrate that one can learn query-specific distance from such sources of information, and the learned distance is more accurate than Euclidean distance and a learned global distance function.

CHAPTER V

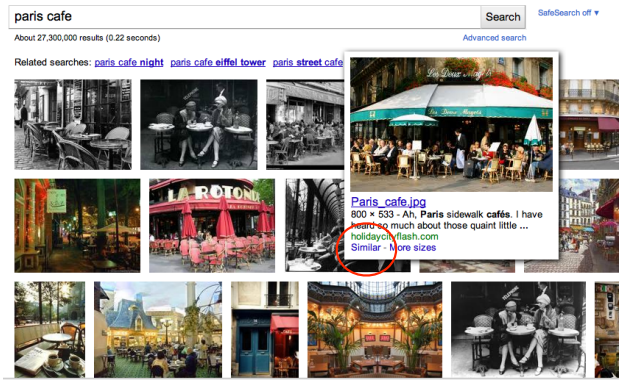
USER STUDIES ON HYBRID IMAGE RETRIEVAL SYSTEMS

Current web image search engines, such as Google or Bing Images, adopt a hybrid search approach in which a text-based query (e.g. "apple") is used to retrieve a set of relevant images, which are then refined by the user (e.g. by re-ranking the retrieved images based on similarity to a selected example). Although chapter 2 and 3 demonstrated that learning from image features and click-patterns can improve the relevance of the top search results and the accuracy of estimated image similarity, it remains to be seen whether adopting such methods in a hybrid image retrieval system can result in measurable improvement in users' efficiency in completing search tasks. This chapter evaluates hybrid image retrieval systems by measuring the improvement in user performance (e.g. time-to-completion) in completing target-search retrieval tasks.

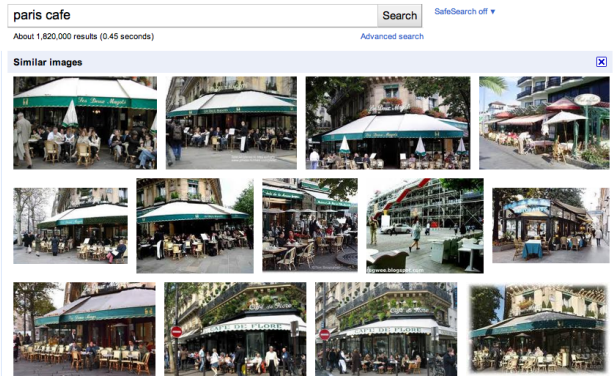
5.1 Introduction

Imagine that you just returned from a trip to Paris and want to write about the famous café located near *Saint-Germain-des-Prés*. Although you do not remember its name, you can still remember the distinctive look of its dome-shaped awning as shown in Figure 35. To find a photo of this café, you can query Web image search engine such as Google or Bing images and browse through the search results until a desired photo is found. If the search

Figure 35: A photo of Café les Duex Magots



(a) step 1 (text-based image retrieval)



(b) step 2 (content-based image re-ranking)

Figure 36: An example of hybrid image retrieval system.

results contain a photo similar to what you are looking for, you can click on the "similar image" link shown below the photo as shown in Figure 36(a). This indicates to the retrieval system that you are looking for other visually similar images and the system will proceed by re-order the search results as shown in Figure 36(b).

Above illustrates an example of a *hybrid* image retrieval system that allows search engine users to enter both text keywords (e.g. the initial text query) and images (e.g. as part of the refinement stage) as queries to describe their search criteria. Figure 37 illustrates the retrieval process of such hybrid image retrieval system: it consists of two parts: the first part retrieves images based on matching keyword query with the meta-data associated with the images, and the second part estimates image similarity and re-ranks images based on user selected exemplar.

The centrality-based ranking and query-specific distance functions proposed in previous

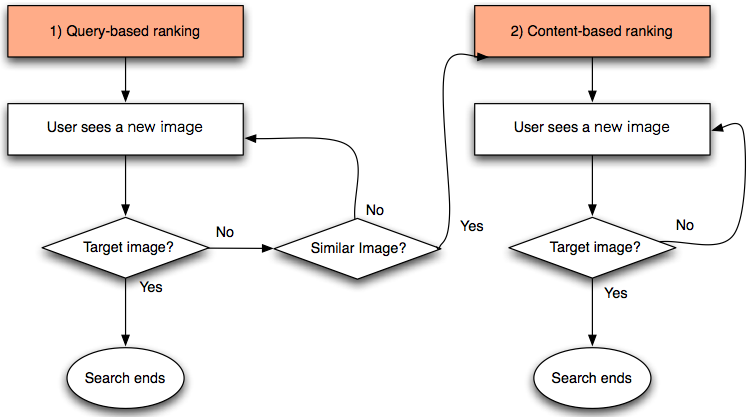


Figure 37: A typical hybrid image retrieval process

Please find this image

The image will disappear in 10 seconds

Figure 38: In target-search experiment, the user is first briefly shown a target image and then instructed to locate the image from an image database using a specific retrieval system.

chapters are designed to improve both the query- and content-based ranking in a hybrid image retrieval system. Although previous experiments in chapter 2 and 3 have demonstrated the benefits of each learning approach in improving the relevance of the search results and the accuracy of image similarity, it remains to be seen whether adopting such methods in a hybrid image retrieval system can result in measurable improvement in users' efficiency in completing a search tasks.

Our goal is to evaluate whether learning to rank from image features and click-patterns made by the users can improve hybrid image retrieval system. Specifically, we propose to evaluate our integrated image retrieval system by conducting a target-search user experiment proposed first by Rodden et al. [80]. In such experiments, the user is first briefly shown

a target image (shown in Figure 38) and then instructed to locate the image from an image database using a specific retrieval system. The effectiveness of the image retrieval systems is measured by average amount of time it takes for the users to successfully complete the target image retrieval task. Comparing with commonly used evaluation approaches such as computing precision and recall scores over an annotated image data-set, such task-based experiments allows one to directly measure the improvement in user performance (e.g. time-to-completion) on actual image retrieval tasks. After all, an automated image retrieval is only meaningful in its service to people [92].

5.2 Large-scale hybrid image retrieval system

We built a large-scale hybrid image retrieval system that supports the retrieval of approximately 250 million Web images. The feasibility of creating such large-scale system relies upon the progress made in three independent areas. First, the advances in large-scale storage and parallel computational infrastructures [24] allow us to compute images features and similarities for significant portion of the Web images. For example, we have access to a parallel computational infrastructure that can evaluate 150 billion similarity comparisons within 7 days. The second advance has been in image representation, including the development of robust image features [62, 73] and robust dimensionality reduction techniques [71, 114, 55, 90, 72, 76] that creates image representations that can be efficiently stored and matched against. Third is the availability of large quantities of user feedback available in anonymized image search engine query logs. Such collective data of individual image search tasks not only reveals relationships between the images and the query, but also amongst the images themselves (e.g., their similarity relationships).

Figure 39 illustrates the five-step process used to build a large-scale hybrid image retrieval system. First, we made a copy of Google image search by caching the search results associated with 300K of the most frequently used keyword queries. Second, we represented each image as a fixed-length feature vector, computed by first concatenating various global features such as color, shape and texture, and then reduce the dimensionality of the feature vector using kPCA with histogram intersection kernel. Third, we extracted image

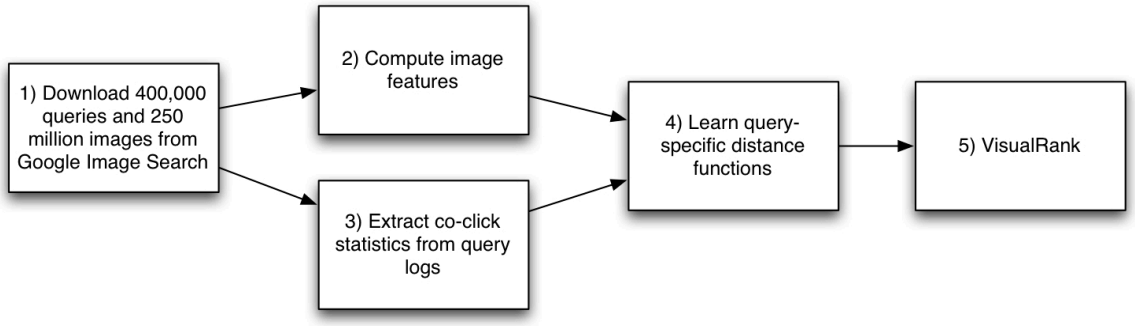


Figure 39: The five-step process used in creating a large-scale integrated hybrid image retrieval system.

co-click statistics from 1-year worth of anonymized Google image search query logs, and selected image triplets as training data using Equation 7 in Chapter 4. Next, we trained query-specific distance functions by adopting large-margin learning approach introduced in Section 3.4, and computed pairwise image similarities for images within the same search results. Given 300K queries and 1000 images retrieved using each query, a total of 300K x 1000 x 1000 / 2 = 150 billion image comparisons are computed using parallel computational infrastructures [24] with 3000 servers. As the last step, we computed image centrality scores with Equation 3 in Chapter 2 using power iteration to re-rank the relevance ordering of the search results.

5.3 Target Search User Experiments

Our goal is to evaluate whether learning to rank from image features and click patterns can improve the effectiveness of a hybrid image retrieval system. Specifically, we adopted the *target-search* experiment methodology [80], designed to simulate the actual retrieval task of locating a specific "target" image using image retrieval systems. The experiment consists of two steps: first, a target image selected from the image database is presented to the subject for a short duration of time; next, the subject is instructed to locate the target image based on their "mental image" using a retrieval or browsing system. The experiments are timed so that the speed of which a task is completed is used to quantify the effectiveness of the retrieval system. This is a variation of simulated work task situation [7], and have been

used previously [18, 68, 15] to evaluate information visualization systems [1].

We adopted target-search experiment methodology for two reasons. First, searching for a target image with a specific criteria in mind is one of the most common mode of image search tasks [23, 16]. For example, in interviews carried out to study how people organize their personal photos, Rodden et al. [81] found that locating a specific image was the most commonly mentioned search task. Jose et al. [53] found that designers searching for photographs reported that they often had a reasonably well defined "mental image" of a photograph that might satisfy their requirement. Therefore, users' efficiency in conducting target-search is a strong indication of the effectiveness of an image retrieval system.

Second, as described in the introduction, hybrid image retrieval systems are suitable for situations where the user does not have access to a readily available query image [2], and has to rely on a set of criteria (such as a "mental image") to guide their search process. Target-search is designed to simulate such search process. Third, as an image retrieval system is only meaningful in its service to people, performance measurement should be anchored in human evaluation, especially when the retrieval system allows users to interact with the search results (e.g. selecting an image exemplar). For example, in a hybrid image retrieval system, the image retrieved depends not only on the ranking function, but also on users' selection of the image exemplars.

We made two choices when designing the study. First, we make the assumption that the search results contain the target image. We conjecture that in practice, if the user cannot find an image in the search results, she or he will formulate another query and repeat this process until the target image is found. Second, in order to better simulate the case where users find images based on a "mental" sketch of the target image, we make the decision to remove the image from the user view after displaying for a short period of time.

[1] Several works have discussed evaluation strategies [61, 13, 68, 51, 118, 14] for information retrieval and information visualization. There are four experimental methodology used in evaluating image retrieval: Annotated datasets [98, 106, 14], simulation experiments [105, 70], user evaluation [93], and click-through analysis [116, 51]. Good performance in target image evaluation would clearly indicate that a system is suited for target-image search applications. This testing methodology is used for mobile image retrieval [3], music retrieval [40].

[2] Otherwise users will opt for content-based image retrieval systems.

5.3.1 Sampling test queries and images

We selected queries belonging to four categories of visual concepts, *person*, *product*, *animals* and *places*. These categories contain many of the most frequent terms people use to query commercial image search engines. Also, as such categories usually have distinctive visual appearances, they are commonly used as a part of benchmark database for evaluation of recognition systems [35, 25].

We also included a fifth category referred as *polysemy*, which are queries with multiple semantic and visual concepts. For example, the query *apple* produces images related to both the company and the fruit. We conjecture that ranking based on image similarity is mostly helpful on the search results retrieved with such ambiguous queries.

We selected 10 queries for each category with the following selection methodology: first, we collected 10,000 of the most frequently searched for queries on Google images during the month of July 2010; second, we uniformly sample queries from this list, and manually assign each query to one of the five categories illustrated above. A query is removed from consideration if it does not fall into any of the five categories, or if the retrieved images contain pornography or other inappropriate content. This process is repeated until each category contains 10 queries. The complete list of queries are shown in Table 6 in Chapter 4. For each query, we collected the top 1000 image using Google image search, with the first 100 images used for training (in combination with their co-click statistics derived from the query logs), and the remaining 900 images for testing.

5.3.2 Experiment User Interface

We follow the grid layout (illustrated in Figure 40) most commonly used in current Web image retrieval systems. The resolution of the viewing area is 1920x1080. As a typical Web image retrieval system displays 20-30 images per page, our system displays 24 images per page. User can use scrollbar to see the next page of images. The system caches all the images in the browser memory at any given time to reduce the display latency. Standard 27-inch monitor will be used, where the browser (chrome) is maximized to occupy the entire screen. When multiple systems are used to conduct studies, the type of monitor, browser,

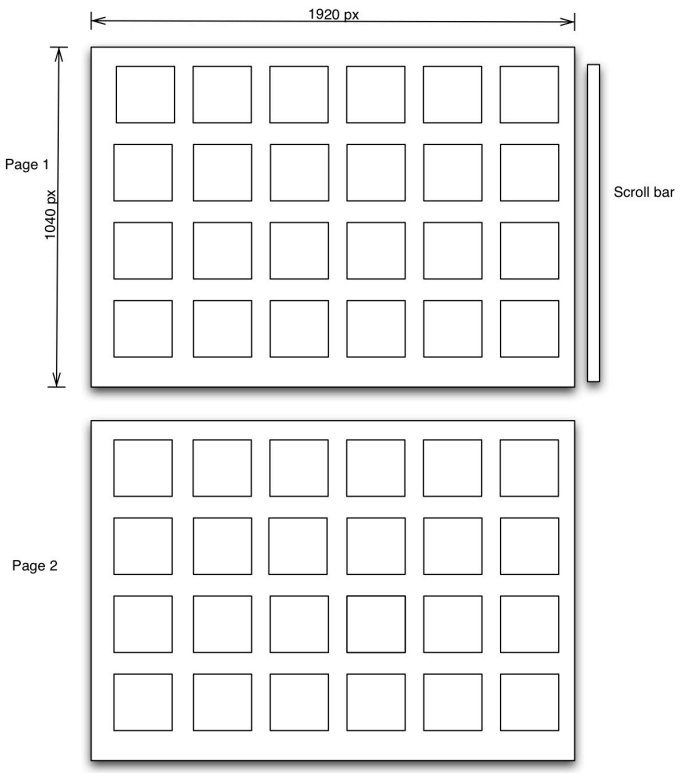


Figure 40: Search results are displayed with grid layout similar to Google and Bing images.

mouse and other interfaces are identical to each other.

Subjects are instructed to use scroll-bar or arrow key to browse through the images. At any time, a subject can switch from text query-based ranking to content-based ranking by selecting an image example. If the clicked image is the target image, then the task is completed. We make the assumption that after an image example is selected, the subject needs to browse through the re-ranked images without making further image selections. This is to simplify our analysis of the retrieval system. In practice, a search engine user may choose to go back to the original search results, or issue a new query if they cannot find the target image in the top re-ranked results. We plan to study such more complex interaction process in future works. We also allow experiment subjects to "skip" any particular task by clicking a button located at the beginning of the search results.

5.3.3 Competing hybrid image retrieval systems

A hybrid image retrieval system shown in Figure 37 consists of two separate ranking functions, and each can affect subjects' efficiency in completing the retrieval tasks. The first part is the computation of the initial ordering of the images (text query-based ranking), and the second part is the ranking of images based on similarity to a selected image exemplar (content-based ranking). Our goal is to measure the effect of our proposed ranking and distance functions on subjects' performance in conducting target-search.

Our study includes two baseline image retrieval systems. The first baseline system is the text query-based image search, denoted as **G**, that does not have the functionality to re-rank images based on image similarity. We use the order in which images are retrieved from Google Images as the output of the baseline text query-based ranking function. Although our primary focus to evaluate the proposed ranking and distance functions presented in previous chapters, it is beneficial to place any potential improvement in comparison with a well-adopted image retrieval method.

The second baseline system, denoted as **GE**, is a hybrid image retrieval system with retrieval process outlined in Figure 37. **GE** has the same text query-based ranking as **G**, but has the additional functionality to re-rank images based on selected exemplar. The pairwise image distance is computed by applying Euclidian distance to image features currently used by Google image search.

We present two competing hybrid image retrieval systems. The first system, denoted as **VE**, replaces the ranking based on the result of Google image search with VisualRank as described in Chapter 2, while maintaining the use of Euclidian distance function to measure image similarity. The second system, denoted as **VQ**, uses query-specific distance function for content-based re-ranking and for the computation of VisualRank. Table 7 presents a summaries of the retrieval systems used in this study.

5.3.4 Experiment procedure

In this experiment, a target-search task is composed of a target image, a text query, a set of images associated with the query, and an image retrieval system. Given the availability

Table 7: Image retrieval systems used in this study.

Retrieval System	Query-based Ranking	Content-based Ranking
G	Google Image Ranking	N/A
GE	Google Image Ranking	Google Euclidean distance
VE	VisualRank	Google Euclidean distance
VQ	VisualRank	Query-specific distance functions

of 44 queries, 10 target images from each query and 4 competing image retrieval systems, there are a total of 1760 different tasks.

Human subjects are asked to complete a set of *randomly* selected tasks within a period of time. The random selection of retrieval tasks ensures that experiments are not biased by the order in which tasks are presented to the subjects – we expect subjects to become familiar with the mechanism of the hybrid retrieval systems and therefore more adept at finding the target image.

15 human subjects participated in this study. They are recruited by a third party who has no knowledge about the goal of the experiments. Each subject conducted 5 hours of experiments, which are divided into 5 segments of 50 minutes task-time followed by 10 minutes rest-time. We do not inform the subjects about the type of ranking algorithm used. This is to prevent subjects from developing strategies that can exploit the artifact of a particular ranking algorithm. At the beginning of the experiment, we present subjects with an instruction page on how to use query-based and hybrid image retrieval system.

Figure 41 presents an example of various steps a subject typically takes to locate the target image. At the beginning of each task, a target image is presented to the subject. After 10 seconds, the target image is removed from his or her view, and the retrieval interface is displayed. We instruct the subjects to examine the search results and click on images that they perceive to be the target image. If the target image is clicked, the system notifies user the task is completed and display the total search time. If the clicked image is not the target image, the retrieval system re-ranks images based similarity to the selected image. Subjects are instructed to search through the re-ranked images until the target image is found. We also provide subjects with an option to "abandon" the task.

Each completed target-search task contains the following information: the ID of the

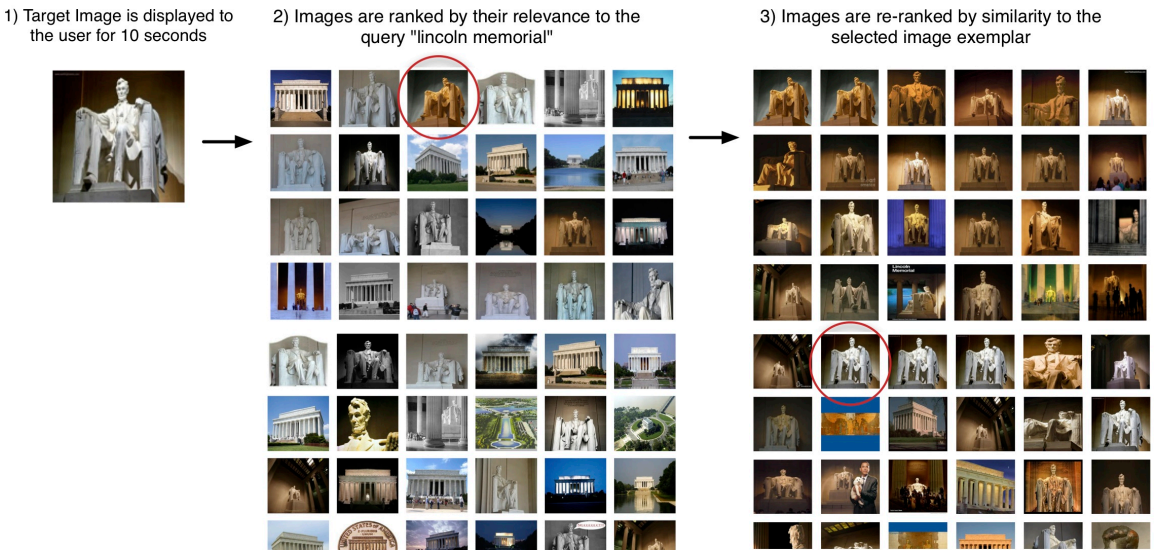


Figure 41: An example of how an experiment subject locates the target image using hybrid image retrieval system.

subject, the type of retrieval system, target image, text query-based ranking, content-based ranking, time-stamped subject interaction with the system. Subjects were told that they were being timed. In order to provide qualitative data, the subjects were also asked to fill in a post-experiment questionnaire.

5.3.5 Evaluation Criteria

We use two quantitative measurements to evaluate the effectiveness of image retrieval systems. The first measurement is **time-to-completion**, the time it takes for the subject to locate the target image. The second is **target-rank**, the position of the target image in the search results, which is closely related to the number of images that users need to examine before the target image is found. Given a text query-based image retrieval system, target-rank is simply the rank of the target image given text query, as we assume that the subjects observe all the images positioned ahead of the target image.

In a *hybrid* image retrieval system, as subjects can change the ordering of the search results by selecting of image exemplar, the value of target-rank depends on the position of the target image given the query, the position of the selected image example, and the position of the target image after re-ranking. Specifically, we define target-rank (TR) for hybrid image retrieval system given the target image t and user selected image exemplar s

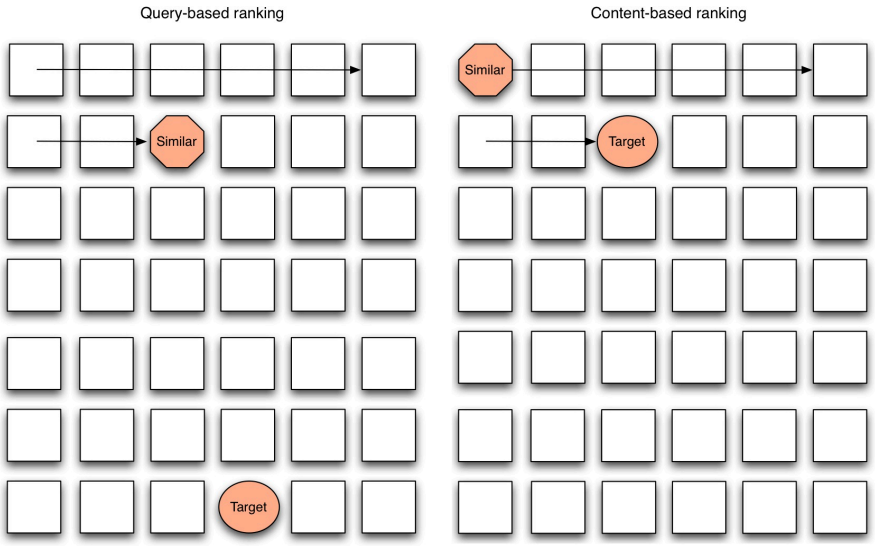


Figure 42: An example of how target-rank is computed for a hybrid image retrieval system: the figure on the left represents the initial position of the target image produced by a text query, and the figure on the right represents the position of the target image after an image exemplar (marked as "similar") is selected. The images covered by the arrows represent those are seen but not selected.

as:

$$TR(t, s) = P(s) + P(t, s) \tag{13}$$

where $P(s)$ is the position of the similar image s in the search results produced by the text query. $P(t, s)$ is the position of the target image t after reordering based on similarity to the image exemplar s. Note that in the initial search results, if the target image is ranked ahead of the best ranked similar image, then target-rank is simply the position of the target image in the initial search results, or $TR(t) = P(t)$.

Figure 42 gives an example of how target-rank is computed for a hybrid image retrieval system: the figure on the left represents the initial position of the target image produced by a text query, and the figure on the right represents the position of the target image after an image exemplar (marked as "similar") is selected. The images covered by the arrows represent those are seen but not selected. As the similar image is at position 9, and the target image is at position 9 after re-ranking, the target-rank is $9 + 9 = 18$. In this case, the target image receives a better target-rank given a hybrid image retrieval system ($AR_q(t, s) = 18$) than a text query-based system ($AR_q(t) = 40$).

Comparing with time-to-completion which is affected by multiple factors not related to the retrieval system (i.e. concentration of the individual when completing that task), target-rank depends only on the ranking functions used and subjects' selection of the similar image.

5.4 Experiment Results

The complete experiment results are shown in Table 8: each row contains a subject identifier, the number of tasks completed and the completion statistics. Column 3-7 displays the average time-to-completion, and column 8-12 display the target-rank of the target images. Due to the random sampling of the tasks, we expect each retrieval system to be used approximately 25% of the time. The retrieval system with the best time-to-completion or target-rank is highlighted.

For example, the first row of Table 8 shows that subject A completed a total of 468 tasks during the experiment, with an average of 34.1 seconds spent on each task using text query-based image retrieval system **G**. The same subject spent on average 28.2, 22.7 and 20.2 seconds on various hybrid image retrieval systems (**GE**, **VE** and **VQ**). This suggests that subject A is most effective at finding the target using **VQ**.

We also computed the average completion statistics across all subjects. Wilcoxon signed rank test is used to measure the statistical significance when two retrieval systems are compared with each other. A value below 0.05 indicates that the comparison is statistically significant. The results shows that most of the comparisons are statistically significant, with the sole exception of comparing average time-to-completion between **VE** and **VQ**.

Figure 43 shows a scatter plot representing the correlation between the time-to-completion and target-rank (with correlation efficient of 0.8204). Each point represents the completion statistics of a subject using a given image retrieval system. The line represents the least square fit of the points. This indicates that time-to-completion and target-rank are highly correlated with each other.

Table 8: Completion statistics for each subject

		Time					**Rank**				
	# of tasks	G	GE	VE	VQ	(var)	G	GE	VE	VQ	(var)
Subject A	368	34.1	28.2	22.7	**20.2**	(12.4)	453	315	195	**172**	(143)
Subject B	324	39.3	31.4	**21.3**	23.4	(14.5)	432	308	**165**	178	(163)
Subject C	334	32.7	19.8	18.6	**17.9**	(11.9)	448	222	202	**185**	(138)
Subject D	401	35.4	27.0	26.4	**24.1**	(13.4)	465	231	212	**201**	(145)
Subject E	348	29.5	24.3	**20.1**	21.0	(14.6)	424	245	207	**198**	(156)
Subject F	322	36.5	20.3	21.7	**18.7**	(13.2)	436	214	193	**178**	(139)
Subject G	402	37.8	28.3	**22.7**	23.7	(11.5)	450	256	201	203	(147)
Subject H	320	41.4	31.4	26.7	**24.5**	(12.8)	471	292	216	**192**	(189)
Subject I	440	36.5	25.3	**23.5**	24.1	(14.2)	483	266	254	**238**	(138)
Subject J	333	34.2	20.4	**18.5**	19.1	(12.9)	420	223	**197**	201	(165)
Subject K	402	41.3	31.0	29.2	**27.5**	(11.2)	482	324	298	**234**	(138)
Subject L	322	48.3	**34.5**	36.9	38.0	(16.9)	429	**234**	243	235	(155)
Subject M	426	28.3	19.5	16.7	**14.5**	(12.9)	481	259	188	**185**	(121)
Subject N	337	35.4	24.0	22.4	**21.4**	(11.1)	455	253	241	**225**	(174)
Subject O	324	36.2	26.2	**24.2**	25.6	(13.2)	465	201	187	**179**	(132)
Average	360	36.1	25.9	23.2	22.7		453	256	213	200	
			GE	VE	VQ			GE	VE	VQ	
Wilcoxon SR		G	**.0000**	**.0000**	**.0000**		G	**.0000**	**.0000**	**.0000**	
		GE		**.0041**	**.0021**		GE		**.0001**	**.0001**	
		VE			.2125		VE			**.0041**	

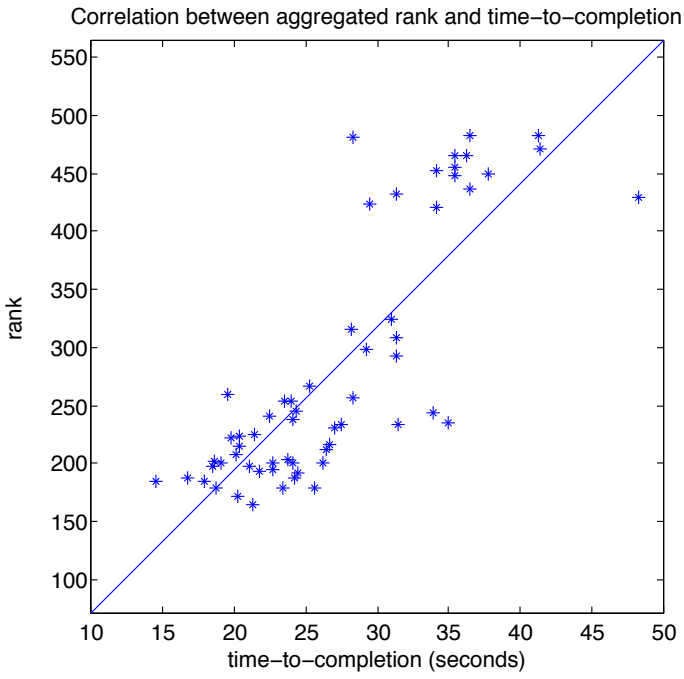


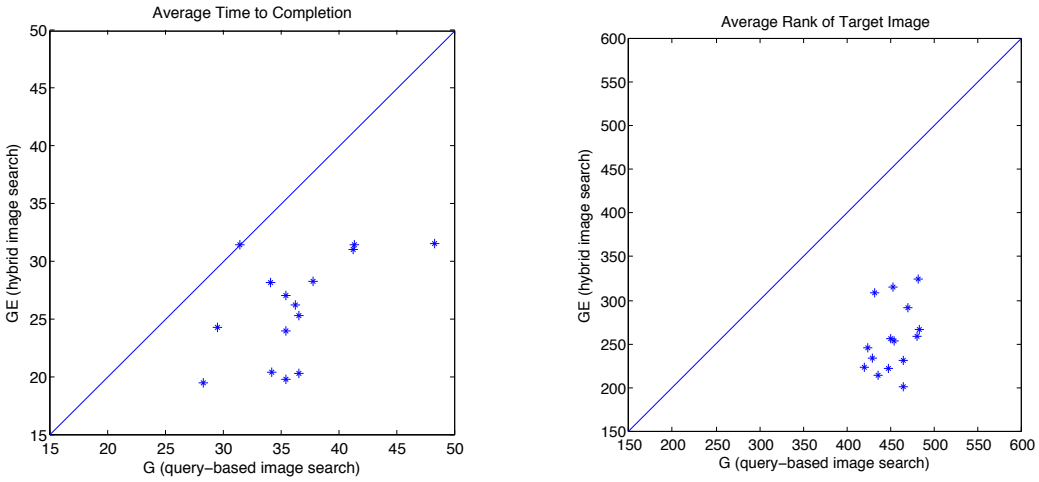
Figure 43: Target-rank and time-to-completion are strongly correlated with each other.

5.4.1 Analysis 1: Hybrid Image Retrieval System

This section compares the two baseline image retrieval systems used in this work: **G** is the text query-based image retrieval system, and **GE** is the hybrid image retrieval system using Euclidian distance to compute pairwise image similarities. Note that these two systems have identical text query-based ranking.

Figure 44 displays the completion statistics of the each subject using scatter plots, with x-axis representing the average completion-time (or target-rank) using **G**, and y-axis representing the average completion-time (or target-rank) using **GE**. Each point on the graph represents the comparison of the performance measurement of a single subject. A point below the diagonal line in Figure 44(a) suggests that a subject is able to find the target image faster using **GE** than using **G**. Similarly, a point below the diagonal line in Figure 44(b) suggests that a subject has to examine more images using **GE** than using **G**.

Figure 44(a) shows that most of the subjects (14 out of 15) are able to complete the target-search task faster using **GE** than using **G**. Similarly, Figure 44(b) shows that all the subjects (15 out of 15) examine fewer images using **GE** than using **G**. This is not



(a) Time to completion

(b) Rank

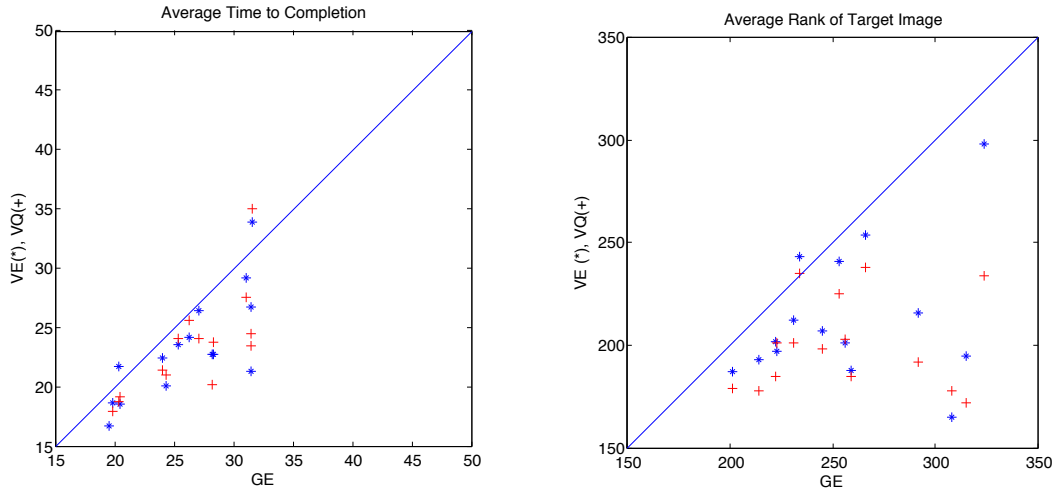
Figure 44: Query-based (G) v.s. Hybrid image retrieval system (GE)

surprising: in a text-query based retrieval system, although images are ranked with respect to the relevance score to the query, in practice it is often difficult to know a user really wants based on a set of keywords, and even more difficult to estimate relevance based on text meta-data associated with the Web images. As a result, subjects often need to browse through large-set of images before the desired image is found. On the other hand, hybrid image retrieval system allows subjects to get to the target image faster by selecting an image example from the top search results.

5.4.2 Analysis 2: VisualRank and Query-specific distance functions

This section evaluates whether learning approaches presented in chapter 2 and 3 can indeed improve users' efficiency in completing target-search tasks.

First, we evaluate the effect of re-ranking based on centrality scores of the images. Specifically we compare **GE**, the baseline hybrid image retrieval system, with **VE**, which re-ranks the search results with VisualRank. Note that these two retrieval systems have identical distance functions for visual re-ranking. Previous experiments (section 2.5) have demonstrated that ranking with image centrality scores can result in fewer irrelevant images among the top search results, we conjecture that the top search results are more likely to contain an image that is perceived as related and visually similar to the target image by the



(a) Time to completion

(b) Rank

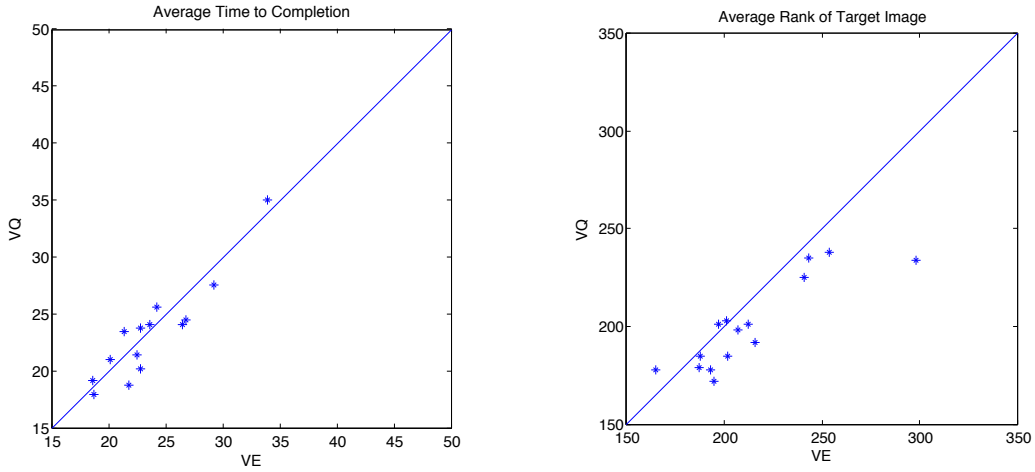
Figure 45: Google Rank (GE) v.s. VisualRank (VE, VQ)

experiment subjects. This will allow experiment subjects to faster locate the target image.

Figure 45 displays the completion statistics of the each subject using scatter plots, with x-axis representing the average completion-time (or target-rank) using **GE**, and y-axis representing the average completion-time (or target-rank) using VisualRank (**GE** and **VQ**). The blue stars (*) represents comparison between **GE** and **VE**, and the red cross (+) represents comparison between **GE** and **VQ**. Similar to Figure 44, each point on the graph represents the performance measurement of a single human subject. A point below the diagonal line in Figure 45(a) suggests that a subject is able to find the target image faster using **VE** (or **VQ**) than using **GE**, and Figure 45(b) suggests that a subject has to examine more images using **VE** (or **VQ**) than using **GE**.

Figure 45 shows that most of the users are able to complete the target-search task faster using either **VE** or **VQ** than using **GE**. The difference is more significant when target-rank is used to compare the two retrieval systems.

Next, we compare **VE**, which uses Euclidean distance to compute image similarities, with **VQ**, which uses query-specific distance functions that are learned from co-click statistics. Note that for the purpose of side-by-side comparison, we used identical query-based



(a) Time to completion

(b) Target Rank

Figure 46: Euclidean distance v.s. Query-specific distance

ranking for **VE** and **VQ** for easy comparisons [3] As experiments presented in section 4.4 demonstrated that a collection of query-specific distance functions can outperform a Euclidean distance (used for all queries) in predicting the outcome of image similarity comparison test, we conjecture that, on average, the target image will be ranked higher in **VQ** than in **VE** given the selected similar image. This will result in lower target-rank, and if the difference is sufficiently large, also result faster time-to-completion.

Figure 46 displays the completion statistics of the each subject using scatter plots, with x-axis representing the average completion-time (or target-rank) using **VE**, and y-axis representing the average completion-time (or target-rank) using **VQ**. Similar to previous scatter plots, , each point on the graph represents the comparison of completion statistics produced by a single human subject. A point below the diagonal line in Figure 46(a) suggests that a subject is able to find the target image faster using **VQ** than using **VE**. Similarly, a point below the diagonal line in Figure 46(b) suggests that a subject has to examine more images using **VE** than using **VQ**.

Figure 46(b) shows that majority of the users (11 out of 15) examined on average fewer images using **VQ** than using **VE**. However, Figure 46(a) shows that the number of

[3] In practice, VisualRank can also be computed using with image similarity computed with query-specific distance functions.

Table 9: The percentage of tasks abandoned by subjects rate when each image retrieval system is used.

Retrieval System	G	GE	VE	VQ
Abandonment Rate	22.5%	17.1 %	12.8 %	9.4%

users completed the tasks using **VE** is about the same those using **VQ**. This suggests that although query-specific distances improves content-based ranking, the difference is not large enough to reduce the target search time.

5.4.3 Analysis 3: Task abandonment

This section analyzes the number of tasks abandoned by the subjects. A task is abandoned when the subject clicks the "skip" button in the option pane. We think abandonment rate offers clear indication on the effectiveness of an image retrieval system. We did not give specific instructions on when the "skip" button can be used. We suspect that subjects are likely to abandon a task when the target photo is difficult to interpret and/or when the subject has experienced frustration in locating the photo in the search results. We aggregated all the abandoned tasks across all subjects and group them based on the type of image retrieval system used.

Table 9 shows the percentage of tasks abandoned given each image retrieval systems, averaged across the tasks for all subjects. It shows that using hybrid image retrieval systems (**GE**, **VE**, **VQ**) resulted in significant lower abandonment rate than text query-based image retrieval system (**G**). It also shows that **VQ** has the lowest abandonment rate with 9.4%. In other words, for every 100 tasks conducted using **VQ**, subjects abandoned on average 9.4 tasks, 3.4 fewer than the next best retrieval system **VE**.

Figure 47 shows the correlation between target-rank and the abandonment rate: 1% of the tasks were abandoned when the target image has a rank of less than 200, 12% abandoned with target-rank of more than 200 but less or equal to 400. The abandonment rate becomes significantly higher when the rank increases 400. 78% of the tasks were abandoned when the target image has the rank of more than 800.

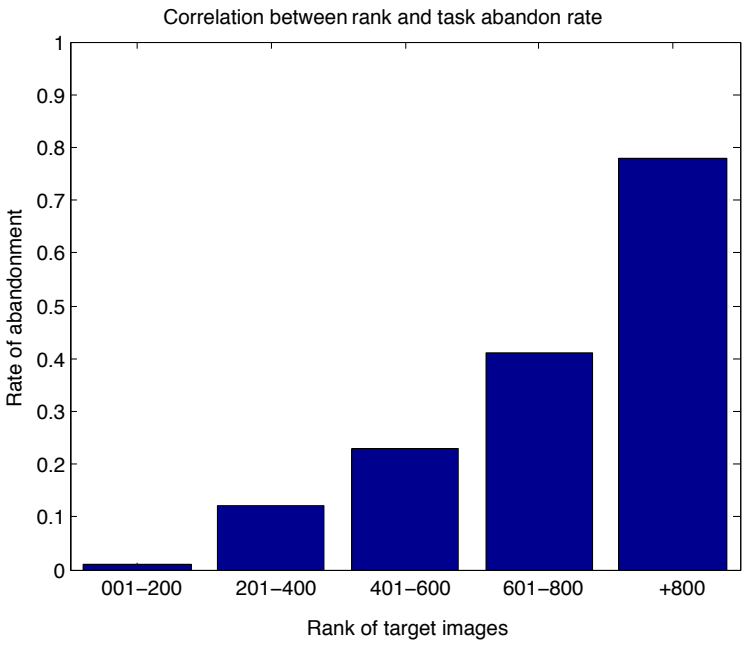


Figure 47: Correlation between target-rank and task abandonment

5.4.4 Analysis 4: Questionnaire

At the end of the experiment, we ask the subjects to rate their experience using the questionnaire shown in Figure 48. The first two questions ask the subjects to rate their previous experiences using Web image retrieval systems, and the next two questions ask them to rate their experience of using hybrid image retrieval system presented in this study.

Since the hybrid image retrieval systems used in this study share identical interfaces and retrieval processes, it is most likely that subjects are not aware of the particular ranking function used for each task. As a result, we do not expect the subjects to provide qualitative comparison of the search results produced by various type of ranking methods. Instead, we ask the subjects to rate their frustration levels when using hybrid image retrieval systems. We also ask the subjects to rate their overall confidence in system's ability to rank images based on similarity to a selected image exemplar.

Figure 49 shows subjects' familiarity with text query-based and hybrid image retrieval systems before the study. it shows that while all the raters use Google images at least once a month, majority of them do not use the functionality of similar image search on a regular basis. The difference is significant: 11 raters use Google image search at least once a week

Participant ID: _____

Target image search experience Questionnaire:
Please tell us about your self and your experience using Similar Images Search.

1. How often do you use Google images? Please select the closest matching answer

Several times a day	Once-a-day	Once-a-week	Once-a-month	Less than Once a month
1	2	3	4	5

2. How often do you use Google **Similar** images? Please select the closest matching answer

Several times a day	Once-a-day	Once-a-week	Once-a-month	Less than Once a month
1	2	3	4	5

3. How often are you frustrated with using the similar image system?

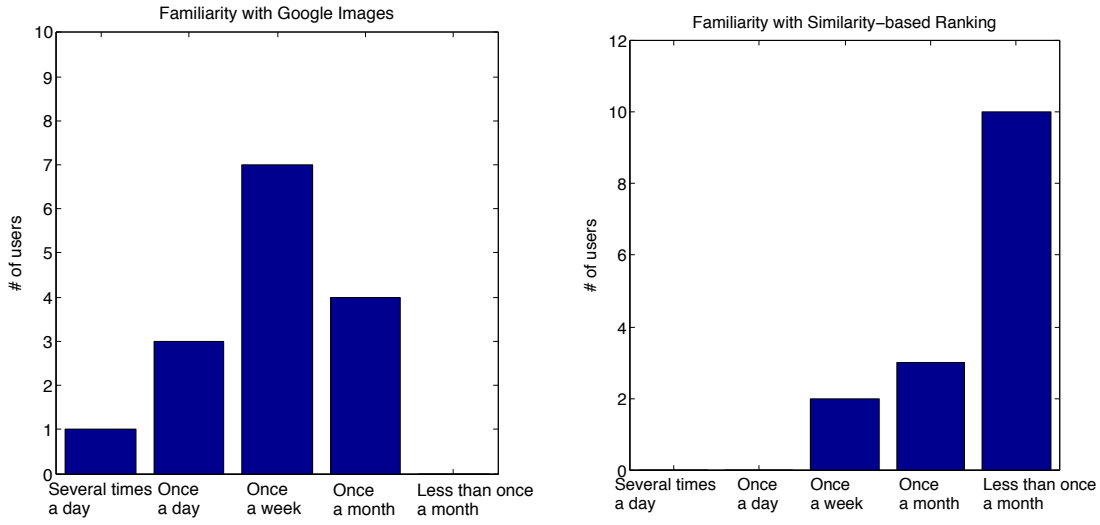
Rarely	Occasionally	Somewhat Often	Often	Very often
1	2	3	4	5

4. How often do you agree that the images that were shown to you after a click were similar to the one that you clicked?

Rarely	Occasionally	Somewhat Often	Often	Very often
1	2	3	4	5

5. Any additional comments you like to share with the researcher

Figure 48: The questionnaires given to the raters.



(a) Answers to question 1

(b) Answers to question 2

Figure 49: Raters' familiarity with Web image retrieval systems.

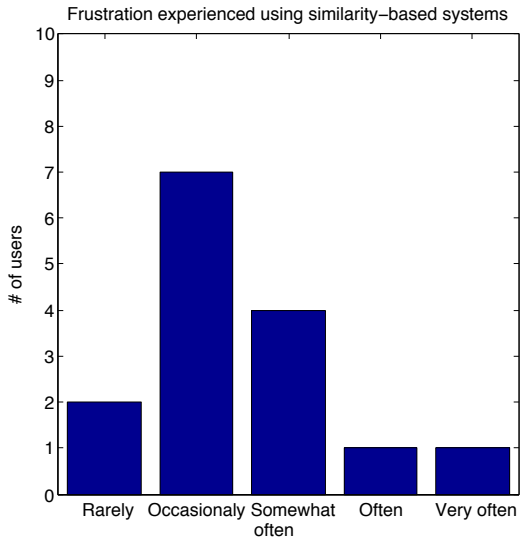
and 4 uses every day; only 2 rater users Google Similar images at least once a week, and majority of them (10 out of 15) uses it less than once a month. This suggests that most raters are not familiar with using image queries to re-rank search results.

Figure 50(a) shows that a significant portion of the subjects (7 out of 15) indicated that they experienced "occasional" frustration when using the hybrid image retrieval systems, followed by "somewhat often." Only a small number of subjects rated "rarely", "often" or "very often." Figure 50(b) shows that significant of subjects (7 out of 15) selected "somewhat often" to describe the frequency in which they agree with the ordering of the images based on image similarity. Also, more subjects chose "often" and "very often" than those chose "occasionally" and "rarely."

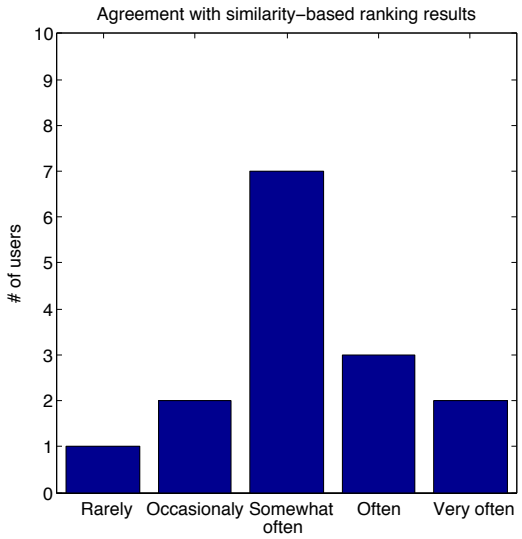
5.5 Conclusions

We presented methods to build an integrated hybrid image retrieval system that covers 400,000 most frequently used queries and up to 250 million Web images. To evaluate various image retrieval system, we conducted a series of user study that measures user efficiency in completing target-search tasks.

The results demonstrates that users achieved consistently lower time to complete the search task on hybrid image retrieval system than text-query based retrieval system, and the



(a) Answers to question 3



(b) Answers to question 4

Figure 50: User satisfaction with hybrid image retrieval system

difference is statistically significant. The results also demonstrate that users using a hybrid image retrieval system that learns to rank using both image features and click-patterns derived from search engine query logs achieved consistently lower completion time than a hybrid image retrieval system which text-query based and content-based ranking decoupled from each other.

The response we received from the post-experiment questionnaire suggests that while majority of the subjects are not familiar with the functionality to search for "similar images," majority of the users are not frustrated.

CHAPTER VI

CONCLUSION

Current Web image search engines, such as Google or Bing Images, adopt a *hybrid* search approach in which a text-based query (e.g. "apple") is used to retrieve a set of relevant images, which are then refined by the user (e.g. by re-ranking the retrieved images based on similarity to a selected example). This approach makes it possible to use both text information (e.g. the initial query) and image features (e.g. as part of the refinement stage) to identify images that are relevant to the user. One limitation of these current systems is that the query- and content-based ranking of images are often computed in a decoupled manner.

This work have proposed two methods to develop an *integrated* hybrid search method which leverages the synergies between query- and content-based image retrieval systems. The first method, presented in Chapter 2, improves the relevance of the query-based search results by computing centrality scores from the visual similarity of the images in the search results. Specifically, we demonstrated that image with higher centrality scores (computed from visual features) are more likely to be considered as relevant to the text query, and such scores can be computed efficiently for large-scale Web search. Also, target-search experiments in Chapter 5 demonstrated that re-ranking using centrality scores can significantly reduces the amount of time it takes for users to find the target image.

The second method, presented in Chapter 3, improves the content-based image ranking by learning query-specific distance functions from the click-patterns made by Web search engine users. Our main intuition was that, for the task of comparing images (retrieved by using query-based search engine), the relative importance of various image features depends on the particular query used. Specifically, we demonstrated that learning query-specific distance functions, using image features and click-patterns of Web search engine users, can more accurately measure image similarity than commonly used Euclidean distance function,

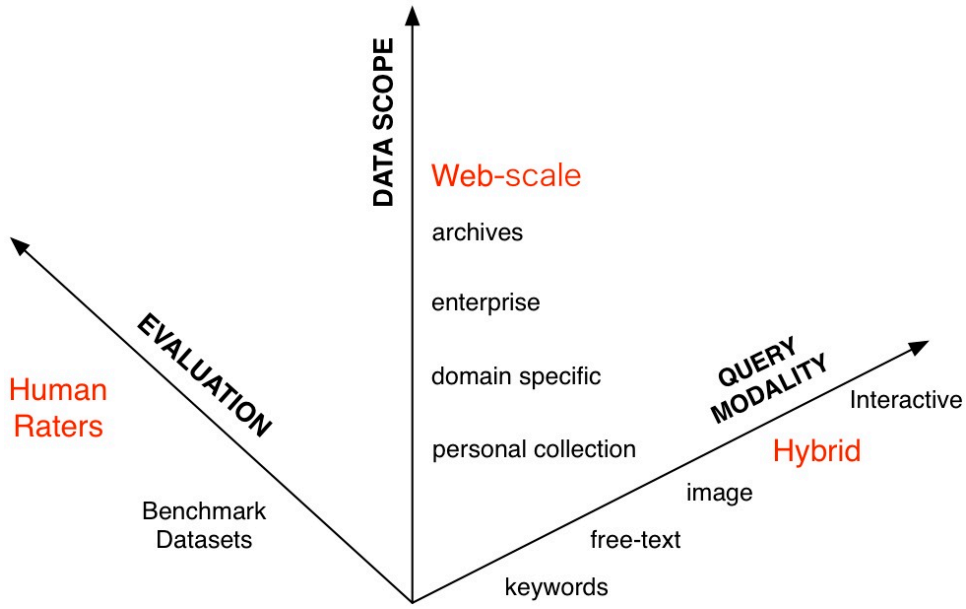


Figure 51: Summary of our work in the context of image retrieval systems.

or a global distance function learned from the same training data.

The learning approaches presented in this thesis share three distinctive characteristics, as illustrated in Figure 51 [1]. First, these approaches use a hybrid modality of information and can only be applied in an *integrated* hybrid image retrieval system as illustrated in Figure 52. For example, VisualRank requires the use of both text (e.g. to generate initial search results) and the image features (e.g. image similarities) to compute centrality scores, and learning query-specific distance functions relies on having access to click-patterns made by users of standard query-based image retrieval system. In other words, while both methods are proposed to improve a separate part of the hybrid image retrieval system, the learning algorithm requires both parts of the system.

The second shared characteristic is that both learning approaches can be efficiently applied to *large-scale* Web image retrieval. For example, Eigen-centrality scores proposed in Chapter 2 can be computed efficiently and in parallel using power iteration methods, the same way PageRank [8] is used for Web search. Similarly, the choice of learning weighted Euclidean distance functions (as opposed to more complex distance functions) over global

[1] The figure is constructed by modifying those used in [23].

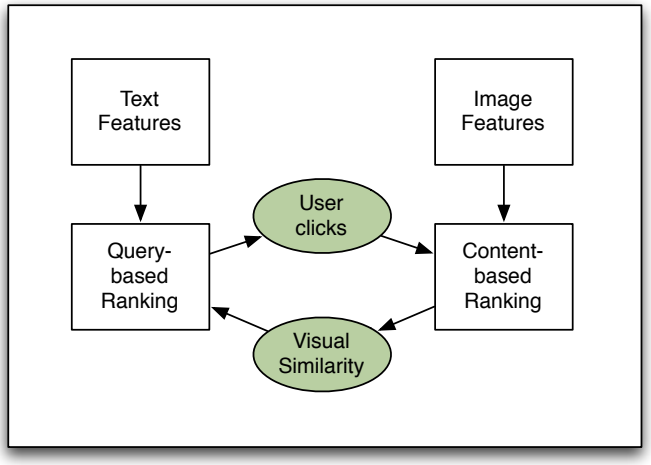


Figure 52: This thesis presents an *integrated* hybrid search method that leverages the synergies between the content- and query-based component of the hybrid image retrieval system.

features allows the computation of image similarity to be done during retrieval-time. In addition to training efficiency, the scalability of a retrieval system is also affected by the cost of obtaining training data. For this reason, our study of learning query-specific distance functions is accompanied by an in-depth analysis of search engine query logs.

The last common characteristic is the prevalence use of *human raters* to measure the performance of image retrieval system. For example, Chapter 2 studies the accuracy of centrality-based ranking by asking users to rate the search results as "irrelevant" to the query. Chapter 4 measured the accuracy of image similarity by conducting perceptual similarity user study and compare the comparison label collected from the users with those derived from co-click statistics and learned distance functions. Chapter 5 measures the performance of hybrid image retrieval system by directly measure the user efficiency in completing retrieval tasks. After all, an automatic image retrieval system is only meaningful in its service to people [93].

6.1 Future directions

There are two potential ways to improve the learning approaches proposed in this work. First, one way to improve query-specific distance function is to allow related text-queries to "share" the learned distance functions. For example, one can first group the text queries

into synsets, and learn *synset-specific* distance functions for each synset. One can derive synsets from expert-knowledge database such as WordNet [66], or from the text and images associated such queries [103]. By allowing training data to be shared among related queries, synset-specific distances can be computed for less commonly used queries. Sharing queries also reduces the number of distance functions that need to be cached by the retrieval system.

Second, as our results have shown that query-specific distance functions can improve ranking accuracy in certain query categories (e.g. polysemy) than others (e.g. animal), the ability to automatically select queries or query categories that are suitable for such distance functions would be beneficial. One possible approach is to measure the disagreement between the co-click statistics and the visual similarity produced by using un-weighted Euclidean distance, and use such disagreement as an indication of whether query-specific distance can be useful.

REFERENCES

[1] "Google similar images."

[2] Ahn, L. V. and Dabbish, L., "Labeling images with a computer game," in *Proc. SIGCHI conference on Human factors in computing systems (CHI)*, (New York, NY, USA), pp. 319–326, ACM, 2004.

[3] Anguera, X., Xu, J., and Oliver, N., "Multimodal photo annotation and retrieval on a mobile phone," in *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, (New York, NY, USA), pp. 188–194, ACM, 2008.

[4] Ashley, J., Flickner, M., Hafner, J., Lee, D., Niblack, W., and Petkovic, D., "The query by image content (qbic) system," *SIGMOD Rec.*, vol. 24, no. 2, p. 475, 1995.

[5] Baluja, S., Seth, R., Siva, D., Jing, Y., Yagnik, J., Kumar, S., Ravichandran, D., and Aly, M., "Video suggestion and discovery for YouTube: Taking random walks through the view graph," in *Proc. 17th International World Wide Web Conference (WWW)*, 2008.

[6] Belongie, S., Malik, J., and Puzicha, J., "Shape matching and object recognition using shape context," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, no. 24, pp. 509–522, 2002.

[7] Borlund, P. and Ingwersen, P., "The development of a method for the evaluation of interactive information retrieval systems," *Journal of Documentation*, vol. 53, pp. 225–250, 1997.

[8] Brin, S. and Page, L., "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, no. 1–7, pp. 107–117, 1998.

[9] Carson, C., Belongie, S., Greenspan, H., and Malik, J., "Blobworld: image segmentation using expectation-maximization and its application to image querying.," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, no. 8, pp. 1026–1038, 2002.

[10] Carterette, B. and Jones, R., "Evaluating search engines by modeling the relationship between relevance and clicks," in *In Proceedings of the Advances in Neural Information Processing Systems (NIPS*, 2007.

[11] Chang, S. K. and Hsu, A., "Image information systems: where do we go from here?," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, no. 5, pp. 431 – 442, 1992.

[12] Chang, S. K., Yan, C. W., Dimitroff, D. C., and Arndt, T., "An intelligent image database system," *IEEE Trans. Softw. Eng.*, vol. 14, pp. 681–688, May 1988.

[13] Chen, C. and Czerwinski, M., "Empirical evaluation of information visualizations: An introduction," *International Journal of Human-Computer Studies*, vol. 5, no. 53, pp. 631–635, 2000.

[14] Clough, P., Müller, H., and Sanderson, M., "Seven Years of Image Retrieval Evaluation," in *ImageCLEF* (Croft, W. B., Müller, H., Clough, P., Deselaers, T., and Caputo, B., eds.), vol. 32 of *The Information Retrieval Series*, pp. 3–18, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[15] Combs, T. T. A. and Bederson, B. B., "Does zooming improve image browsing?," in *Proceedings of the fourth ACM conference on Digital libraries*, DL '99, (New York, NY, USA), pp. 130–137, ACM, 1999.

[16] Conniss, L R, A. J. A. and Graham, M. E., "Information seeking behaviour in image retrieval: Visor i final report," tech. rep., Institute for Image Data Research (Library and Information Commission Research Report 95)., 2000.

[17] Cox, I. J., Miller, M. L., Minka, T. P., Papathomas, T. V., and Yianilos, P. N., "The bayesian image retrieval system, pichunter: Theory, implementation, and psychophysical experiments," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 9, no. 1, pp. 20–37, 2000.

[18] Cox, I. J., Miller, M. L., Omohundro, S. M., and Yianilos, P. N., "Target testing and the pichunter bayesian multimedia retrieval system," in *Proceedings of the 3rd International Forum on Research and Technology Advances in Digital Libraries*, (Washington, DC, USA), pp. 66–, IEEE Computer Society, 1996.

[19] Cox, T. F. and Cox, M. A. A., *Multidimensional Scaling*. Chapman and Hall, 1994.

[20] Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B., "An experimental comparison of click position-bias models," in *Proceedings of the international conference on Web search and web data mining*, WSDM '08, (New York, NY, USA), pp. 87–94, ACM, 2008.

[21] Crucianu, M., Ferecatu, M., and Boujemaa, N., "Relevance feedback for image retrieval: a short survey," in *In State of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction including Datamodels and Languages (DELOS2 Report*, 2004.

[22] Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S., "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Symposium on Computational Geometry (SCG)*, pp. 253–262, 2004.

[23] Datta, R., Joshi, D., Li, J., and Wang, J., "Image retrieval: Ideas, influences, and trends of the new age.," *ACM Computing Surveys*, no. 2, 2008.

[24] Dean, J. and Ghemawat, S., "MapReduce: Simplified data processing on large clusters," in *Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI-04)*, (San Francisco, California), pp. 137–150, 2004.

[25] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[26] Fergus, R., Perona, P., , and Zisserman, A., "Object class recognition by unsupervised scale-invariant learning," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 264–271, 2003.

[27] Fergus, R., Perona, P., and Zisserman, A., "A visual category filter for Google images," in *Proc. 8th European Conference on Computer Vision (ECCV)*, pp. 242–256, 2004.

[28] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P., "Query by image and video content: the QBIC system," *Computer*, vol. 28, no. 9, pp. 23–32, 1995.

[29] Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T., "Evaluating implicit measures to improve web search," *ACM Transactions on Information Systems*, vol. 23, p. 2005, 2005.

[30] Frey, B. J. and Dueck, D., "Clustering by passing messages between data points," *Science*, pp. 972–976, 2007.

[31] Friedman, N., Geiger, D., and Goldszmidt, M., "Bayesian network classifiers," *Machine Learning*, pp. 131–163, 1997.

[32] Frome, A., Singer, Y., Sha, F., and Malik, J., "Learning globally-consistent local distance functions for shape-based image retrieval and classification," in *Proc. 11th IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

[33] Frome, A., Singer, Y., Sha, F., and Malik, J., "Learning globally-consistent local distance functions for shape-based image retrieval and classification," *ICCV*, 2007.

[34] Grauman, K. and Darrell, T., "The pyramid match kernel: Efficient learning with sets of features," *Journal of Machine Learning Research*, vol. 8, pp. 725–760, April 2006.

[35] Griffin, G., Holub, A., and Perona, P., "Caltech-256 object category dataset," Tech. Rep. 7694, California Institute of Technology, 2007.

[36] Griffin, G., Holub, A., and Perona, P., "Caltech-256 object category dataset," tech. rep., Caltech, 2007.

[37] Gupta, A. and Jain, R., "Visual information retrieval," *Commun. ACM*, vol. 40, pp. 70–79, May 1997.

[38] Haveliwala, T., "Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search," *IEEE Transactions on knowledge and Data Engineering*, no. 4, pp. 784–796, 2003.

[39] Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J., "An algorithmic framework for performing collaborative filtering," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, (New York, NY, USA), pp. 230–237, ACM, 1999.

[40] Hoashi, K., Hamawaki, S., Ishizaki, H., Takishima, Y., and Katto, J., "Usability evaluation of visualization interfaces for content-based music retrieval systems."

[41] Hoi, S. C. H., Liu, W., Lyu, M. R., and ying Ma, W., "Learning distance metrics with contextual constraints for image retrieval," in *Proc. Computer Vision and Pattern Recognition*, pp. 2072–2078, Murray Hill, 2006.

[42] Huijsmans, D. P. and Sebe, N., "How to complete performance graphs in content-based image retrieval: Add generality and normalize scope," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 245–251, February 2005.

[43] Indyk, P., "Stable distributions, pseudorandom generators, embeddings, and data stream computation," in *Proc. 41st IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 189–197, 2000.

[44] Indyk, P., Motwani, R., Raghavan, P., and Vempala, S., "Approximate nearest neighbor–towards removing the curse of dimensionality," in *Proc. 30th ACM Symp. on Computational Theory*, pp. 604–613, 1998.

[45] Jain, V. and Varma, M., "Learning to re-rank: Query-dependent image re-ranking using click data," in *Proceedings of the International World Wide Web Conference*, March 2011.

[46] Jing, Y. and Baluja, S., "Visualrank: Applying pagerank to large-scale image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, pp. 1877–1890, November 2008.

[47] Jing, Y., Baluja, S., and Rowley, H., "Canonical image selection from the web," in *Proc. 6th International Conference on Image and Video Retrieval (CIVR)*, pp. 280–287, 2007.

[48] Jing, Y., Rowley, H., Rosenberg, C., Wang, J., Tsai, D., and Covell, M., "Google image swirl, a large-scale content-based image browsing engine," in *Submission to World Wide Web (WWW, 2012)*, 2012.

[49] Jing, Y., Rowley, H., Wang, J., Rosenberg, C., Tsai, D., and Covell, M., "Google image swirl," in *(In submission) World Wide Web (WWW)*, 2012.

[50] Jing, Y., Pavlović, V., and Rehg, J. M., "Boosted bayesian network classifiers," *Mach. Learn.*, vol. 73, pp. 155–184, November 2008.

[51] Joachims, T., "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, (New York, NY, USA), pp. 133–142, ACM, 2002.

[52] Joachims, T., Granka, L., Inc, G., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G., "Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search," *ACM Transactions on Information Science (TOIS*, 2007.

[53] Jose, J. M. and Fkner, J., "Spatial querying for image retrieval: a user-oriented evaluation," pp. 232–240, ACM, 1998.

[54] Joshi, D., Wang, J. Z., and Li, J., "The story picturing engine–a system for automatic text illustration," *ACM Transactions on Multimedia, Computing, Communications and Applications*, no. 1, pp. 68–89, 2006.

[55] Ke, Y. and Sukthankar, R., "Pca-sift: A more distinctive representation for local image descriptors," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 506–513, 2004.

[56] Kennedy, L. and Naaman, M., "Generating diverse and representative image search results for landmarks," in *Prof. 17th International World Wide Web Conference (WWW)*, pp. 297–306, 2008.

[57] Kleinberg, J. M., "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, no. 5, pp. 604–632, 1999.

[58] Kondor, R. I. and Lafferty, J., "Diffusion kernels on graphs and other discrete structures," in *Proc. 19th International Conference on Machine Learning (ICML)*, pp. 315–322, 2002.

[59] Kulis, B., Jain, P., and Grauman, K., "Fast similarity search for learned metrics," vol. 31, pp. 2143–2157, December 2009.

[60] Lazebnik, S., Schmid, C., and Ponce, J., "A sparse texture representation using affine-invariant regions," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 319–324, 2003.

[61] Lesk, M. and Salton, G., "Relevance assessments and retrieval system evaluation," *Information Storage and Retieval*, vol. 3, pp. 241–247, 1968.

[62] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, no. 2, pp. 91–110, 2004.

[63] Ma, W.-Y. and Manjunath, B. S., "A toolbox for navigating large image databases," *Multimedia System*, no. 7, pp. 184–198, 1999.

[64] Manning, C. D., Raghavan, P., and Schütze, H., *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008.

[65] Mikolajczyk, K. and Schmid, C., "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, no. 10, pp. 1615–1630, 2005.

[66] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K., "Wordnet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, pp. 235–244, 1990.

[67] Moghaddam, B., Tian, Q., Lesh, N., Shen, C., and Huang, T. S., "Visualization and user-modeling for browsing personal photo libraries," *International Journal of Computer Vision*, vol. 56, p. 2004, 2004.

[68] Müller, H., Müller, W., Squire, D. M., Marchand-Maillet, S., and Pun, T., "Performance evaluation in content-based image retrieval: overview and proposals," *Pattern Recogn. Lett.*, vol. 22, pp. 593–601, April 2001.

[69] Neumann, D. and Gegenfurtner, K. R., "Image retrieval and perceptual similarity," *ACM Trans. Appl. Percept.*, vol. 3, pp. 31–47, January 2006.

[70] Nguyen, G. P. and Worring, M., "Interactive access to large image collections using similarity-based visualization," *J. Vis. Lang. Comput.*, vol. 19, pp. 203–224, April 2008.

[71] Nistér, D. and Stewénius, H., "Scalable recognition with a vocabulary tree," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2161–2168, 2006.

[72] Nowak, E. and Jurie, F., "Learning visual similarity measures for comparing never seen objects," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[73] Oliva, A. and Torralba, A., "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, May 2001.

[74] Palmer, S., Rosch, E., and Chase, P., "Canonical perspective and the perception of objects," *J. Long and A. Baddely, Attention and Performance IX*.

[75] Pentland, A., Picard, R., and Sclaroff, S., "Content-based manipulation of image databases," *International Journal of Computer Vision (IJCV)*, no. 3, pp. 233–254, 1996.

[76] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A., "Object retrieval with large vocabularies and fast spatial matching," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[77] Platt, J. C., *Fast training of support vector machines using sequential minimal optimization*, pp. 185–208. Cambridge, MA, USA: MIT Press, 1999.

[78] Radlinski, F., Kurup, M., and Joachims, T., "How does clickthrough data reflect retrieval quality?," in *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, (New York, NY, USA), pp. 43–52, ACM, 2008.

[79] Robertson, S., Vojnovic, M., and Weber, I., "Rethinking the esp game," in *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, CHI EA '09, (New York, NY, USA), pp. 3937–3942, ACM, 2009.

[80] Rodden, K., Basalaj, W., Sinclair, D., and Wood, K., "Evaluating a visualisation of image similarity (poster abstract)," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, (New York, NY, USA), pp. 275–276, ACM, 1999.

[81] Rodden, K. and Wood, K. R., "How do people manage their digital photographs?," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '03, (New York, NY, USA), pp. 409–416, ACM, 2003.

[82] Rogowitz, B. E., Frese, T., Smith, J., Bouman, C. A., and Kalin, E., "Perceptual image similarity experiments," in *In SPIE Conference on Human Vision and Electronic Imaging*, pp. 576–590, 1998.

[83] Rosch, E., "Natural categories," *Cognitive Psychology*, vol. 7, pp. 573–605, 1973.

[84] Roweis, S. T. and Saul, L. K., "Nonlinear dimensionality reduction by locally linear embedding," *SCIENCE*, vol. 290, pp. 2323–2326, 2000.

[85] Rubner, Y., Tomasi, C., and Guibas, L. J., "The earth mover's distance as a metric for image retrieval," *IJCV*, 2000.

[86] Ruthven, I. and Lalmas, M., "A survey on the use of relevance feedback for information access systems," *Knowledge Engineering Review*, vol. 18, no. 1, 2003.

[87] Salton, G. and McGill, M. J., *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill Book Co., 1983.

[88] Salton, G., "The state of retrieval system evaluation," *Inf. Process. Manage.*, vol. 28, pp. 441–449, March 1992.

[89] Saraiva, P. C., Silva de Moura, E., Ziviani, N., Meira, W., Fonseca, R., and Riberio-Neto, B., "Rank-preserving two-level caching for scalable search engines," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, (New York, NY, USA), pp. 51–58, ACM, 2001.

[90] Schindler, G., Brown, M., and Szeliski, R., "City-scale location recognition," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[91] Schultz, M. and Joachims, T., "Learning a distance metric from relative comparisons," in *Proc. 16th Conference on Advances in Neural Information Processing Systems (NIPS)*.

[92] Shirahatti, N. V. and Barnard, K., "Evaluating image retrieval," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, (Washington, DC, USA), pp. 955–961, IEEE Computer Society, 2005.

[93] Shirahatti, N. V. and Barnard, K., "Evaluating image retrieval," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 955–961, IEEE Computer Society, 2005.

[94] Simon, I., Snavely, N., and Seitz, S. M., "Scene summarization for online image collections," in *Proc. 12th International Conference on Computer Vision (ICCV)*, 2007.

[95] Simon, I., Snavely, N., and Seitz, S. M., "Scene summarization for online image collections," in *Proc. 11th International Conf. on Computer Vision (ICCV)*, 2007.

[96] Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R., "Content-based image retrieval at the end of the early years.," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, no. 12, pp. 1349–1380, 2000.

[97] Smith, J. and Chang, S. F., "Visualseek: a fully automated content-based image query system," in *Proc. 4th ACM international conference on Multimedia*, (New York, NY, USA), pp. 87–98, ACM, 1996.

[98] Smith, J. R., "Image retrieval evaluation," in *Proceedings of the IEEE Workshop on Content - Based Access of Image and Video Libraries*, CBAIVL '98, (Washington, DC, USA), pp. 112–, IEEE Computer Society, 1998.

[99] S.Shalev-Schwartz, Y.Singer, and N.Srebro, "Pegasos: primal estimated sub-gradient solver for svm," *ICML*, 2007.

[100] Swain, M. J. and Ballard, D. H., "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.

[101] T. Papathomas, I. Cox, P. Y. M. M. T. M. T. C. J. G., "Psychophysical experiments on the pichunter image retrieval system," *Journal of Electronic Imaging*, vol. 10, pp. 170–180, 2001.

[102] Tenenbaum, J. B., Silva, V., and Langford, J. C., "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, Dec. 2000.

[103] Tsai, D., Jing, Y., Liu, Y., A.Rowley, H., Ioffe, S., and M.Rehg, J., "Large-scale image annotation using visual synset," *ICCV*, 2011.

[104] Uchihashi, S. and Kanade, T., "Content-free image retrieval by combinations of keywords and user feedbacks," *Proc. International conference on Image and Video Retrieval (CIVR)*, 2005.

[105] Vendrig, J., Worring, M., and Smeulders, A. W. M., "Filter image browsing - exploiting interaction in image retrieval," in *Proceedings of the Third International Conference on Visual Information and Information Systems*, VISUAL '99, (London, UK), pp. 147–154, Springer-Verlag, 1999.

[106] Vogel, J. and Schiele, B., "Performance evaluation and optimization for content-based image retrieval," *Pattern Recogn.*, vol. 39, pp. 897–909, May 2006.

[107] W. H. Hsu, L. K. and Chang, S., "Video search reranking through random walk over document-level context graph," in *Proc. 15th International Conference on Multimedia*, pp. 971–980, 2007.

[108] Weinberger, K., Blitzer, J., and Saul, L., "Distance metric learning for large margin nearest neighbor classification," pp. 1437–1480, 2006.

[109] Winder, S. and Brown, M., "Learning local image descriptors," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[110] Wu, G. and Chang, E. Y., "Formulating context-dependent similarity functions," in *In ACM International Conference on Multimedia (MM)*, pp. 725–734, ACM Press, 2005.

[111] X. He, W.-Y. M. and Zhang, H., "Imagerank: spectral techniques for structural analysis of image database," in *Proc. International Conference on Multimedia and Expo*, pp. 25–28, 2002.

[112] X. Zhu, Z. G. and Lafferty, J. D., "Semi-supervised learning using gaussian fields and harmonic functions," in *Proc. 20th International Conference on Machine Learning (ICML)*, pp. 912–919, 2003.

[113] Xing, E., Ng, A., Jordan, M., and Russel, S., "Distance metric learning, with applications to clustering with side-information," in *Proc. 15th Conference on Advances in Neural Information Processing Systems (NIPS)*, pp. 450–459, 2002.

[114] Y. Ke, R. S. and Huston, L., "Efficient near-duplicate detection and sub-image retrieval," in *Proc. ACM International Conference on Multimedia (ACM MM)*, pp. 869–876, 2004.

[115] Yang, L. and Jin, R., "An efficient algorithm for local distance metric learning," in *in Proceedings of AAAI*, 2006.

[116] Zamir, O. and Etzioni, O., "Grouper: a dynamic clustering interface to Web search results," *Computer Networks*, vol. 31, pp. 1361–1374, May 1999.

[117] Zhang, H., Berg, A. C., Maire, M., and Malik, J., "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," in *In CVPR*, pp. 2126–2136, 2006.

[118] Zhang, J., *Visualization for Information Retrieval*. John Wiley and Sons, 2008.

[119] Zhou, X. and Huang, T., "Relevance feedback for image retrieval – a comprehensive survey," *Multimedia Systems*, 2004.